2020 24th International Conference Information Visualisation (IV) | 978-1-7281-9134-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/IV51561.2020.00083

Using Machine Learning to Explore the Relation Between Student Engagement and Student Performance

Fidelia Orji, Julita Vassileva
Computer Science Department
University of Saskatchewan
Saskatoon, Canada
fidelia.orji@usask.ca, jiv@cs.usask.ca

***Abstract*—Engagement in learning activities is an important factor that affects student performance in education. According to research, student engagement involves the degree of passion, interest and attention that they exhibit in their educational environment. In the traditional learning system, educators encourage students to engage in their learning activities through various teaching strategies such as making them pay attention, take notes, ask questions and participate actively in the learning processes. Sometimes, educators call on a specific student to answer a question as a means of encouraging the student to participate in learning processes. Nowadays, engagement strategies for learning are changing, especially with the use of technology-enhanced learning systems (TELS) in education. As a result, improving the engagement level of students in online learning environments remains an open research question that needs to be explored. This research is part of a preliminary study on discovering ways of increasing student engagement in an online learning system through data-driven interventions. Student engagement in this research is determined using objective data (activity logs of a specific undergraduate course in a TELS). Activity log is unbiased data and a reflection of students' actual learning behaviours (uncontrolled). In this study, we mined the log of students' learning activities from a TELS used for an undergraduate course to explore differences between students' learning behaviours as they relate to their engagement level and academic performance (measured in terms of final grade points in a course). We employed supervised (Random Forest) and unsupervised (Clustering) machine learning approaches in exploring the relations. The approaches identified an interesting pattern on student engagement and show that engagement and assessment scores are good predictors of student academic performance. Assessment scores are measured with results of quizzes and assignments performed by the students in the TELS, while academic performance is measured with the final grade of the student in the course. The implications of our findings are discussed**.

Keywords—student engagement, student performance, machine learning, supervised and unsupervised machine learning, clustering, random forest, educational data mining, learning pattern, online learning, academic performance, technology-enhanced learning systems.

I. Introduction

Increasing use of online learning systems nowadays for both eLearning courses and blended learning (a combination of face-to-face teaching with web-based TELS) has resulted in the generation of a huge volume of learning data. Research in learning analytics is harnessing the data to understand the real learning behaviour of students and determine factors that improve learning success. Increasing attention is paid to student engagement in learning as one of the factors affecting student performance. Various research studies revealed the importance of student engagement in both face-to-face teaching [1] and online learning systems [2]. The resulting theoretical models on student engagement especially in higher education have formed the basis for discussion about the relation between learning engagement and other learning factors such as student performance.

Previous studies investigating student engagement and performance in learning usually take the form of surveys in measuring engagement. In survey-based studies, students typically answer questions designed using existing models for student engagement. This approach uses self-reports, which are often biased and subjective, hence the results may not be realistic. However, student engagement for learning could be determined using objective data such as logged students' activities in learning-related tasks. According to research, mining data of students' learning logs could reveal their real learning behaviour which may help in identifying patterns of learning that are successful [3]. Thus, analyzing learning systems data will help in determining students' actual learning behaviours and in reporting their learning progress which will assist educators in their decision-making process. The analysis could also support assessing the relation between learning variables (for example student engagement) and student performance.

Previous research pointed to the need to use data from the students' learning behaviours and characteristics in predicting their performance [4], especially in TELS. In this study, we mined students' learning logs to gain insights about the relationship between their engagement level and their academic performance. We used a dataset collected from a blended learning system used in a large first-year undergraduate class at our University. The dataset provided information on students' activities and interactions with the TELS. We performed an exploratory analysis and applied unsupervised and supervised machine learning methods to students' activities to determine how they affect their academic performance.

This research seeks to find the relation between engagement variables, segment the variables and assessment scores using cluster analysis, explore the characteristics of the different segments, and predict student performance using the engagement variables and assessment scores.

2375-0138/20/$31.00 ©2020 IEEE
DOI 10.1109/IV51561.2020.00083

Specifically, the goal of this study is to answer the following specific research questions:

RQ1: How do the student engagement variables relate to each other? Are there identifiable groups of students with certain patterns of engagement variables values that perform better?

RQ2: What is the relationship between student engagement variables and their assessment scores?

RQ3: What is the relationship between student engagement variables, assessment scores and actual academic performance?

This research adds to the existing studies on the use of learning analytics in understanding students learning progress and in supporting educational institutions in making appropriate decisions that will improve students' learning. It also adds to existing research on student engagement in higher education with new insight obtained from learning from log data.

II. BACKGROUND

A. Student Engagement for Learning

According to student involvement theory for higher education, the learning and personal growth of students in an educational program increase as the quality and quantity of students' involvement increase [5]. The theory postulates that involvement could be a quantitative measure such as time spent on learning activities or qualitative such as measures of learning goals. The measures could be general or specific (that is involving entire student experience in learning or just experience in a specific course). Based on the theory, highly-involved students devote considerable energy and time in studying and participating in academic activities. Moreover, research suggested that universities that highly engage their students with a variety of relevant learning activities that help to improve the learning outcome of their students may be considered to have a higher learning quality than a university with less engaging activities for students' learning [6]. This is because the more students study, practice, perform assessments and get feedback, the deeper their understanding of what they are learning.

Research has studied student engagement in education in various forms. For instance, Pace [7] one of the earliest researchers on student engagement developed the College Student Experience Questionnaire (CSEQ) tool. Pace reported that students that devoted more time and energy to learning tasks gained a lot from their studies in terms of college experience and application of concepts learned to concrete situations. There is growing importance in understanding the effect of student engagement on their learning experience and to institutions of learning. Various communities such as Community College Survey of Student Engagement (CCSSE) and National Survey of Student Engagement (NSSE) have been developed for assessing the quality of effort and participation of students in useful learning activities. In line with the relevancy of student engagement, research highlighted the role of institutions in improving engagement as it affects institutions' and students' performance [6].

As online education continues to penetrate both blended and distance learning systems, the need for improving students' learning experience and performance in online systems become a vital issue to explore. Various researchers have studied student learning experience and engagement using different survey-based approaches. For example, Delfino [1] in a survey-based study investigated factors affecting student engagement and its association with academic performance using statistical methods. However, few studies exist on exploring student engagement using their actual learning behaviour in an uncontrolled learning system (a system where students can log in and study to meet their set learning goals at will). Thus, this research studies student engagement using their actual learning activities logs and machine learning approach.

B. Machine Learning Algorithms and Educational Datasets

Intelligent educational systems learn from student activities interaction data and adapt/improve/personalize their strategies and content. They use data mining techniques from supervised and unsupervised learning algorithms. The educational data mining (EDM) area has a more general focus; it explores datasets generated from students' learning activities using different machine learning and data mining algorithms to understand students' learning processes and their learning environment [8]. With the help of the algorithms, researchers have been able to find answers to specific problems concerning students' learning experience and effectiveness. For instance, in identifying students that are likely to fail a particular course using the students' previous performance data and decision tree algorithm, a predictive model was built using engineering students' data [9]. The model was used to detect in advance the students that are likely to fail a course so that adequate assistance for improvement of their learning could be provided for them. On the other hand, in predicting students that will likely proceed to pursue a postgraduate degree, a study collected data from senior undergraduate students with the use of questionnaire and applied decision tree algorithm in Weka, the result showed a classification accuracy of 88% [10].

Studies have made efforts to analyze the learning interaction of students in various systems to obtain insights concerning different students' learning approaches and to answer some research questions based on specific goals. Some of the studies try to model students based on their learning behaviour. For example, Amershi et al. [11] built a framework with both supervised and unsupervised classification algorithms for identifying useful learning interaction of students. The framework was applied to two different environments of learning using logged and eye-tracking data. The authors suggested that their framework could be used for automatic classification of learning behaviours of new students on online learning systems. Many other works demonstrate how artificial intelligence techniques and statistical tools can be applied in evaluating and adapting e-learning systems to students [12]. For example, the usage patterns of learners on the e-learning system can be classified according to usage level for the purpose of adapting the content and structure of the e-learning system and also for detecting learners that are not regular.

Most higher institutions of learning use course management and e-learning systems for posting and providing access to course materials for students. According to research, these systems do not offer educators the opportunity to evaluate learning processes and course effectiveness based on activities performed by students [13]. Thus, several studies providing insight from educational data through the use of clustering algorithms have been performed. Parack et al. [14] in a study on profiling and grouping students based on their academic records, applied apriori algorithm to students' academic records to extract association rule for profiling and the k-means algorithm was used in grouping the students based on their learning pattern. They reported that their implemented algorithms could provide an efficient way of profiling students. Similarly, research on improving accessibility of learning objects through a personalized learning setting proposed a combination of k-means algorithm and self-organizing map for clustering and ranking learning objects [15]. Furthermore, the Expectation-Maximization (EM) clustering algorithm is frequently used for the clustering of data in machine learning. The algorithm has been applied to educational data for various purposes. For instance, research has shown that the application of EM to course evaluation data discovered useful student profiles [16]. Bogarin et al. [17] proposed a model that first applied the EM algorithm to group students on basis of their performance and based on the result of the clustering, students' behaviour for each cluster was discovered. A review of various applications of clustering to educational datasets for different purposes is provided in [18].

III. METHODOLOGY

We performed this study to identify different groups of students with important characteristics related to their learning and to predict student performance using objective data. To answer the research questions, we performed some exploratory analysis and report our results over the same set of features.

A. Data Collection and Processing

The data used for this research was collected in a blended learning course (Biology) taken by undergraduates in a Canadian university. The students involved in the course used a TELS called MindTap system [19]. The system logged data on students' actions, activities, and assessments.

To obtain some relevant features that might assist us in determining engagement level and assessment scores of students, we cleaned and prepared the dataset using Python. We removed some features that might not be relevant to our analysis. Furthermore, students' records without logged activities and actions (null data) were deleted. After the data preprocessing, we were left with data (records of students' activities) from four hundred and ninety (490) students. Some of the features selected from the dataset for the analysis include the following:

Total time spent in MindTap (TimeOnTask) – This feature shows the total amount of time that each student has spent in MindTap on various activities such as Homework, Assignments, Quizzes, and Readings. The time was logged in hours, minutes and seconds. We converted the total time to minutes as there was nothing logged on seconds.

Number of logins (NumberofLogins) – This displays the total number of logins in MindTap for each student.

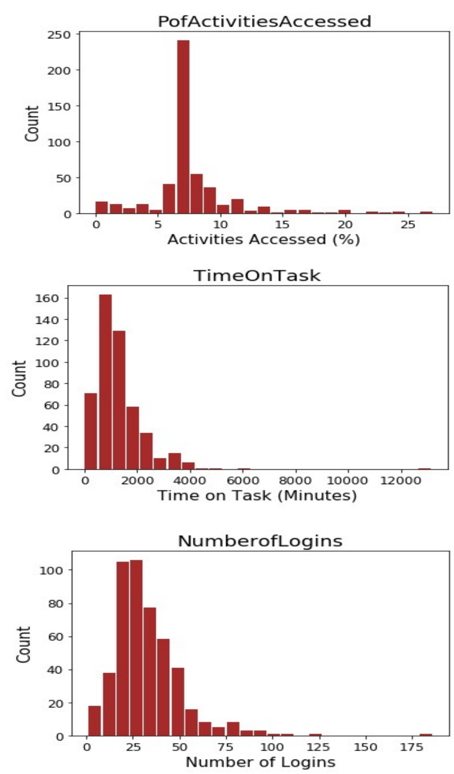
Percentage of Activities Accessed (ActivitiesAccessed) – This indicates the percentage of activities accessed by each student out of the total number of activities assigned.

Overall Score in percentage (AveAssessmentScore) – It indicates the average performance score for each student based on the score of all relevant assessments performed on the MindTap system.

Furthermore, we explored the dataset to get information on the distribution of values within each of the selected features. Table 1 and Figure 1 gives information on the description and distribution of the features on our dataset. Table 1 shows the total number of student records on the dataset as 490 and other statistics about each feature. For example, the mean of NumberofLogions is 32.5, the minimum is 1.0, the maximum is 186,0 and the 25th, 50th, 75th percentiles are 21, 29, and 40 respectively. Figure 1 helps us to determine whether the distribution of values within the features are different.

TABLE 1. SUMMARY STATISTICS OF OUR DATASET

	PofActivitiesAccessed	TimeOnTask	NumberofLogins	AveAssessmentScore
count	490.000000	490.000000	490.000000	490.000000
mean	7.663265	1286.955102	32.522449	78.693367
std	3.512595	1001.846133	18.580519	22.026185
min	0.000000	0.000000	1.000000	0.000000
25%	7.000000	706.250000	21.000000	71.150000
50%	7.000000	1101.000000	29.000000	86.030000
75%	8.000000	1613.000000	40.000000	94.287500
max	27.000000	13150.000000	186.000000	100.000000



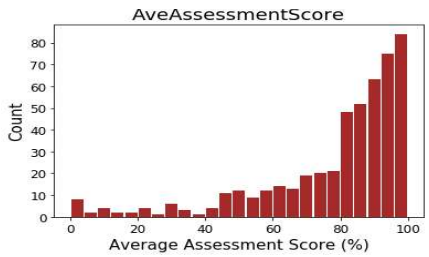


Fig. 1. Distribution of each feature in our dataset

As can be seen from Figure 1, two of the features NumberofLogins and TimeOnTask are skewed right (their tails extend towards the right). The figure shows that the two features contain outliers. The feature ActivitiesAccessed is roughly symmetric. The assessment performance AveAssessmentScore is left-skewed. The figures show that the distribution of values within each feature is different.

Based on the result of our dataset exploration, the outliers were deleted and a total of four hundred and eighty-eight (488) students' records were used for the analysis. Also, approaches that will optimize the distribution of the dataset features were chosen for the analysis.

B. *Data Analysis*

To determine the degree of association between the selected engagement variables (learning activities features), we performed a correlation analysis to measure the relationship between the engagement variables using the Spearman correlation coefficient in Python. The result is shown in Table 2.

In determining different groups of students based on their engagement variables, we applied clustering, an unsupervised machine learning method suitable for partitioning data meaningfully to discover hidden patterns in it. The clustering used the Expectation-Maximization (EM) [20] algorithm as implemented in Weka. The algorithm uses a random initialization and iterative process which alternates the expectation, *E* and maximization, *M* steps continuously until the algorithm convergence [21]. It tries to optimize the parameters of the model to best explain the dataset through the maximization of the likelihood of the data in the final clusters. Research has shown that the EM algorithm is useful when using a real-world dataset that involves clustering small scenes (features) where k-means cannot perform well [22]. Several studies that proposed students' modelling and profiling via a data-driven approach have used the EM algorithm in achieving various goals concerning students learning [16], [17]. The algorithm instead of trying to maximize the difference in mean of data instances maximizes the likelihood of a given data in the final cluster using computation of the likelihood of cluster membership based on probability distribution. The algorithm has the advantage of approximating the observed distributions of features according to mixtures of different distributions in the clusters and it automatically determines the appropriate number of clusters. This process of hyperparameter tuning of the algorithm helps in determining the optimal number of clusters for a given clustering problem. The result of the clustering is shown in Table 3.

Predicting Academic Performance of Students

Having gained insight on the relationship between engagement variables and assessment performance through unsupervised machine learning approach – clustering, we decided to investigate the degree of association between engagement variables and academic performance (final grade in a course) of students. We employed a supervised machine learning algorithm called random forest in investigating the impact of engagement and assessment scores on academic performance. The random forest algorithm is a good option when features in a dataset are not well scaled. It performs classification and regression tasks. For this study, we applied the random forest algorithm for a regression task. The algorithm is very stable and it has reduced bias because it combines multiple decision trees through an ensemble learning method and builds trees using random data points from the training set. The ensemble learning uses bagging technique and this allows individual decision trees (subsets) to run in parallel without interacting with each other. The algorithm uses the average outcome of each tree in predicting its final outcome and this helps to improve its prediction performance and prevents overfitting through random sampling of data subsets.

Using Scikit-Learn implementation of the random forest algorithm in Python, we constructed a model that can predict students' academic performance in a university course based on their engagement variables and assessment scores. We applied percentage split technique to our dataset, 80% for training and 20% as a test set. To find the number of trees parameter value that can best predict academic performance, we performed hyperparameter tuning. The number of trees parameter was optimized based on root mean squared error (RMSE). The parameter values tested were 10, 20, 30, 40, 50, 60, 100, 200, 500, and 1000. We obtained optimal parameters setting when the number of trees parameter was set to 40, the random state to 42 and the other parameters used their default settings. The model was then evaluated using the test set to determine how it will perform on a new dataset. The result of the prediction is presented in the next section.

IV. RESULTS AND DISCUSSION

A. *The Relation between Engagement Variables*

The results of the Spearman correlation in Table 2 show that the three engagement variables used in this research: ActivitiesAccessed, TimeOnTask, and NumberofLogins have a positive correlation among them. The positive correlation indicates that the variables will likely perform well as engagement measures. This answers our research question on the relation between the engagement variables.

TABLE 2. SPEARMAN CORRELATION RESULT FOR ENGAGEMENT VARIABLES

	ActivitiesAccessed	**TimeOnTask**	**NumberofLogins**
ActivitiesAccessed	1.0000	0.5246	0.4533
TimeOnTask	0.5246	1.0000	0.6002
NumberofLogins	0.4533	0.6002	1.0000

B. The Relationship between Engagement Variables and Assessment Scores

The application of clustering to our dataset identified interesting students' categories as clusters. Each of the clusters significantly differs in their characteristics as shown in Table 3. The three clusters created were labelled as C0 for the first cluster, C1 and C2 for the second and third clusters respectively. Students grouped in C0 (148 students) were highly engaged as shown by the measures of engagement variables (ActivitiesAccessed, TimeOnTask, and NumberofLogins) and they had an excellent performance (89.561) as indicated in their assessment measure. The students in this group are assumed to have adopted a dedicated approach to learning which consequently affected their assessment performance. For C1, the students in this group (116 students) were not actively engaged as indicated in their engagement measures. They did not show much commitment to their learning activities and it affected their assessment performance (56.104). The students grouped in cluster C2 (224 students) were more committed to their learning activities and they performed better than those in C1.

In answering one of our research questions, we can say that the students in cluster C0 performed better than those in the other two groups. This means that the higher the engagement for learning activities, the better the assessment scores. This result is consistent with other studies in literature that revealed that student performance relates to their level of engagement [23]. Moreover, the result shows that the C0 group that was highly engaged performed better in assessments than the others who were not deeply engaged.

TABLE 3. CLUSTERING RESULTS OF THE EM ALGORITHM

	Clusters		
Features	**C0 (Mean)**	**C1 (Mean)**	**C2 (Mean)**
ActivitiesAccessed	9.891	5.644	7.000
TimeOnTask	1886.597	654.602	1128.790
NumberofLogins	45.114	18.838	30.035
AveAssessmentScore	89.561	53.413	85.948

The number of students in each cluster is as follows: C0 has 148 students (30%), C1 contains 116 (24%), and C2 contains 224 (46%).

C. The Relationship between Engagement Variables, Assessment Scores and Actual Academic Performance

The result of our random forest model shows that there is some relation between our selected features (engagement variables and assessment scores) and the students' actual academic performance. The evaluation result of the model on the test set shows an accuracy of 84.10% and root mean square error (RMSE) of 12.35. Accuracy was calculated using the mean absolute percentage error.

To determine the usefulness of each feature in improving the model, we checked the relative importance of the features using Scikit-Learn. The result shows features importance as follows: AveAssessmentScore contributes 60%, TimeOnTask contributes 20%, NumberofLogins contributes 13%, and ActivitiesAccessed contributes 7%. The assessment scores (AveAssessmentScore) is the highest contributing factor, followed by time on task (TimeOnTask) and the percentage of activities accessed (ActivitiesAccessed) is the least.

D. The Implications of our Results

Mining learning logs of students' activities could provide useful information for profiling and grouping them based on their learning patterns. Research has shown that students have different learning characteristics that affect their ability to learn. Thus, grouping students with similar engagement levels will provide an interesting way of tailoring learning interventions to students based on their engagement needs. Appropriate interventions optimizing learning of the different levels could be provided. Such intervention might involve the use of both internal and external motivators such as visualizations, incentive mechanisms or persuasive technology in encouraging students to actively participate in their learning activities. These approaches could be applied in TELS using learning data as they have been shown to improve participation. For example, research has shown that presenting different levels of contribution of users in an online community using visualization has a significant effect on improving participation [24]. Consequently, automating the grouping on technology-enhanced learning systems (TELS) using clustering model as shown in this research, and reporting the data using visualizations that educators understand, will help in providing useful information on the progress of learners. The information will assist educators in determining if the students are deeply involved in their learning activities. If it is found that the students are not committed to their learning as they should, the influencing factors (such as design, structure and pedagogical elements of the TELS) could be investigated and this will assist institutions in taking proper decisions on improving students learning experience and performance.

Our prediction model in this research has shown that engagement levels and assessment scores of students in TELS are good predictors of their academic performance. With the use of this model in TELS, individual students can be presented information on how their study practices and assessments affect their performance and this will increase their awareness of what their final grade will be if they do not improve in their study practices. Moreover, the model will assist educators in identifying on time students that are likely to fail/drop (at-risk students) a course. Hence, appropriate measures for helping at-risk students could be initiated automatically without much resources from educators which will help to save resources for other purposes. According to research, improving student engagement could help educational institutions in addressing problems of high dropout rate, low performance and boredom among students [25].

V. CONCLUSION

Student engagement as a vital construct in understanding student learning behaviour could be used in evaluating technology-enhanced learning systems on their ability to properly impact students' learning especially now that higher education institutions incorporate TELS as part of the required learning medium for students. The data from these systems provide information on how the students engage with them to achieve their learning goals. Analysis of the data provides educators with reliable information on students' learning progress which will help them in identifying students learning needs and in making decisions on how to improve the learning experience of students.

This paper presented preliminary work on students' group modeling based on their learning interaction to gain an understanding of how their engagement indicators on TELS affect their academic performance. It applied machine learning methods to educational data obtained in a blended learning environment to achieve its goal. The work highlighted the relationships between engagement level and student academic performance and how machine learning algorithms could help educators in monitoring and responding to students' learning progress issues automatically, thereby allowing them to spend their time on other pedagogical issues.

Higher education institutions could apply the group modeling approach in this research in detecting how effective a TELS is at inspiring students for learning and also in offering automatic adaptive interventions based on this group modeling which might be difficult to accomplish for individual students (using the predictive model). The adaptivity of the systems will be in response to observed pattern of learning needs.

REFERENCES

[1] A. P. Delfino, "Student engagement and academic performance of students of Partido State University," *Asian J. Univ. Educ.*, vol. 15, no. 1, pp. 22–41, 2019.

[2] H. J. Kim, A. J. Hong, and H. D. Song, "The roles of academic engagement and digital readiness in students' achievements in university e-learning environments," *Int. J. Educ. Technol. High. Educ.*, vol. 16, no. 1, p. 21, Dec. 2019.

[3] G. McCalla, "The Ecological Approach to the Design of E-Learning Environments: Purpose-based Capture and Use of Information About Learners," *J. Interact. Media Educ.*, vol. 2004, no. 1, p. 3, May 2004.

[4] M. Vahdat, A. Ghio, L. Oneto, D. Anguita, M. Funk, and M. Rauterberg, "Advances in Learning Analytics and Educational Data Mining," in *23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015 - Proceedings*, 2015, pp. 297–306.

[5] A. W. Astin, "Student involvement: A developmental theory for higher education," *J. Coll. Stud. Dev.*, vol. 40(5), pp. 518–529, 1999.

[6] G. D. Kuh, "The national survey of student engagement: Conceptual and empirical foundations," *New Dir. Institutional Res.*, vol. 2009, no. 141, pp. 5–20, 2009.

[7] C. R. Pace, "Measuring the quality of college student experiences: An account of the development and use of the college student experiences questionnaire," *High. Educ. Res. Inst.*, pp. 1–136, 1984.

[8] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 40, no. 6. pp. 601–618, Nov-2010.

[9] B. R. S. Kabra R R, "Performance Prediction of Engineering Students using Decision Trees," *Int. J. Comput. Appl. (0975 - 8887) Vol. 36- No.11, December 2011*, vol. 36, no. 11, 2011.

[10] V. P. Breşfelean, "Analysis and predictions on students' behavior using decision trees in weka environment," in *Proceedings of the International Conference on Information Technology Interfaces, ITI*, 2007, pp. 51–56.

[11] S. Amershi and C. C. Conati, "Combining Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments," *JEDM-Journal Educ. Data Min.*, vol. 1, no. 1, pp. 1–54, Nov. 2009.

[12] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in *Proceedings of the 20th International Conference on Very Large Data Base*, 1994.

[13] M. E. Zorrilla, E. Menasalvas, D. Marín, E. Mora, and J. Segovia, "Web usage mining project for improving Web-based learning sites," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2005, vol. 3643 LNCS, pp. 205–210.

[14] S. Parack, Z. Zahid, and F. Merchant, "Application of data mining in educational databases for predicting academic trends and patterns," in *Proceedings - 2012 IEEE International Conference on Technology Enhanced Education, ICTEE 2012*, 2012.

[15] A. S. Sabitha and D. Mehrotra, "User centric retrieval of learning objects in LMS," in *Proceedings of the 2012 3rd International Conference on Computer and Communication Technology, ICCCT 2012*, 2012, pp. 14–19.

[16] E. Trandafili, A. Allkoçi, E. Kajo, and A. Xhuvani, "Discovery and evaluation of student's profiles with machine learning," in *ACM International Conference Proceeding Series*, 2012, pp. 174–179.

[17] A. Bogarín, C. Romero, R. Cerezo, and M. Sánchez-Santillán, "Clustering for improving Educational process mining," in *ACM International Conference Proceeding Series*, 2014, pp. 11–15.

[18] A. Dutt, "Clustering Algorithms Applied in Educational Data Mining," *Int. J. Inf. Electron. Eng.*, 2015.

[19] "MindTap - The leading digital learning tool – Cengage." [Online]. Available: https://www.cengage.com/mindtap/. [Accessed: 07-Aug-2020].

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin, " Maximum Likelihood from Incomplete Data Via the EM Algorithm ," *J. R. Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1–22, Sep. 1977.

[21] G. Celeux and G. Govaert, "A classification EM algorithm for clustering and two stochastic versions," *Comput. Stat. Data Anal.*, vol. 14, no. 3, pp. 315–332, Oct. 1992.

[22] N. Sharma, A. Bajpai, and R. Litoriya, "Comparison the various clustering algorithms of weka tools," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 5, pp. 73–80, 2012.

[23] H. Lei, Y. Cui, and W. Zhou, "Relationships between student engagement and academic achievement: A meta-analysis," *Soc. Behav. Pers.*, vol. 46, no. 3, pp. 517–528, 2018.

[24] J. Vassileva and L. Sun, "Evolving a Social Visualization Design Aimed At Increasing Participation in a Class-Based Online Community," *Int. J. Coop. Inf. Syst.*, vol. 17, no. 04, pp. 443–466, Dec. 2008.

[25] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Review of Educational Research*, vol. 74, no. 1. pp. 59–109, 2004.