IJITGEB

INTERNATIONAL JOURNAL IN INFORMATION TECHNOLOGY IN GOVERNANCE, EDUCATION AND BUSINESS
Vol. 3, No. 1, 2021
ISSN 2686-0694 (Print)
e-ISSN 2721-0030 (Online)

Assessing Online Learners' Access Patterns and Performance Using Data Mining Techniques

Myra Collado Almodiel
myra.almodiel@upou.edu.ph
University of the Philippines Open University (UPOU), Philippines

Abstract: In an online learning environment where students do not have direct face-to-face interactions with instructors, observing their learning behaviors is quite challenging. As we move from traditional to virtual classrooms, it is important to look into the learning methods and approaches that can be used to have a deeper understanding of student participation in an online learning environment, especially in the Philippine setting where this type of research is relatively very scarce or not investigated at all. This study showed the advances in employing the combination of cluster analysis and data mining in assessing the students' online access patterns and performance. Using the data mining process, this study analyzed the access patterns and performance of undergraduate students in an online course in an open university in the Philippines using the data generated from an open-source LMS called Moodle. Moreover, descriptive and inferential statistics, clustering, and visualization techniques were employed to identify and analyze the students' behavioral patterns and preferences based on the login frequency, frequency of accessing course materials, number of views and posts in the collaborative learning logs and discussion, announcements, and self-introduction forums, and frequency of submission and completion of assignments, exams, and projects. Results of the study suggest that students are more engaged if given more activities and opportunities for collaboration such as in discussion forums. Results also showed that a big part of students' access to the course site is to access (view and post) the discussion forums and view the introduction forum where they learn something from their classmates. The data mining techniques, particularly the statistics, clustering approach, and visualization can be a useful tool in analyzing the online learners' access, patterns, and performance.

Keywords: Data Mining, Learning Management System (LMS), online learning, student performance, clustering

INTRODUCTION

In an online learning environment where students do not have direct face-to-face interactions with teachers, observing students' learning behavior is quite a challenge. Considering the huge amount of data on students' activities that an online learning management system collects and stores every day, analyzing the students' performance can be a time-consuming and challenging task for a teacher. Assessment of the students' performance in complex computer-supported collaborative learning or inquiry-based learning scripts is a tiresome and time-consuming process for the teachers, who should take into consideration a huge amount of parameters (Dimopoulos et al., 2013).

Recently, there is a growing number of research studies on the use of data mining techniques on educational data to analyze the performance of the students. Studies suggest that data mining is very useful in examining students' learning behavior in an online learning environment due to its potential in "analyzing and uncovering the hidden information of the data itself which is hard and very time consuming if done manually" (Mohamad & Tasir, 2013). The increasing interest in the field of data mining in educational systems is making educational data mining "a new growing research community" (Romero & Ventura, 2007).

Data Mining Tools and Techniques

Generally speaking, data mining techniques are powerful tools for discovering hidden knowledge and have great potential to reveal online learning behavioral patterns, preferences, progress, and more (Hung & Zhang, 2008). Data mining techniques can be used to determine the learning behavioral patterns such as students' participation behavior, learning style, and preferences on course resources. For instance, Hung & Zhang (2008) were able to identify students' behavioral patterns and preferences in the online learning processes, differentiated active and passive learners, and found important parameters for performance prediction, using clustering and decision tree analysis. Alfiani & Wulandari (2015) focused on mapping students using the K-mean Cluster algorithm to reveal the hidden pattern and classify students based on their demographic (sex, origin, GPA, grade of certain courses), and the average of course attending. On the other hand, Park et al. (2016), used the class analysis and clustering approach to analyze blended learning courses by online behavior data.

In a survey on educational data mining from 1995 to 2005, Romero & Ventura (2007) identified several EDM techniques :

1. Statistics and Visualizations;
2. Web mining (Clustering, classification, and outlier detection; association rule mining and pattern mining); and
3. Text Mining.

In the study on "Tools for Educational Data Mining" (Slater, et al., 2017) several EDM tools which are frequently used by researchers to conduct EDM analyses were reported. These were summarized in Table 1.

Table 1

General Purpose Tools for Educational Data Mining

EDM Tools	Functions	Sample Tools
Data manipulation	For manipulation, cleaning, and formatting of data, as well as for feature engineering and data creation	Microsoft Excel Google Sheets EDM Workbench Python Jupyter notebook SQL
Algorithmic analysis	To model and predict processes and relationships in educational data	RapidMiner WEKA SPSS KNIME Orange KEEL Spark MLlib
Visualization	To build interactive visual interfaces for gaining knowledge and insight from data, as well as communicating important implications for learning to students and teachers	Tableau D3.js

Data Mining Process

According to Romero, et al. (2007), the data mining process in e-learning consists of four steps:

1. Collect data from an LMS system on students' usage and interaction.
2. Preprocess the data, meaning the data is cleaned and transformed into an appropriate format to be mined.
3. Apply data mining data algorithms to build and execute the model that discovers and summarizes the knowledge of interest for the user (teacher, student, administrator, etc.)., where
4. Interpret, evaluate and deploy the results.

Access patterns and student performance

The relationship between online access patterns and student performance is a subject that captures the interest of many researchers (Butrous, 2011). By identifying patterns of learning behavior, teachers may be able to “identify those students who require additional assistance and intervention” (Haig & Falkner, 2013). Previous studies on the relationship between online access patterns and student performance have shown that there is a strong positive relationship between access patterns and students’ performance. (Butrous, 2011).

A research study by Haig & Falkner (2013) showed that “student’s pattern of study, as measured by LMS usage, correlates with their final grade in a course”. In the same research, it was reported that “students who are more engaged with the course perform better in terms of their final grade” (Haig & Falkner, 2013). Bagarinao (2011), in his study on "Learners’ Access Patterns and Performance in an Online Course in Science, Technology, and Society" revealed that “learners who have visited the site and read messages more frequently got higher scores in the examinations administered for the course”.

This research study aims to analyze the undergraduate students' access patterns and performance in an online learning environment using data mining techniques using the data generated from an open-source LMS called Moodle, using selected standard report plugins. A combination of statistical analysis, visualization, and clustering approach was used to analyze the students' access patterns and performance.

RESEARCH DESIGN & METHODS

Unit of Analysis

The data used in this study contained the log records of 45 undergraduate students in a three-unit introductory course on information technology in an open university in the Philippines extracted from the course site using Moodle (Modular Object-Oriented Dynamic Learning Environment) an open-source learning management system. Those who dropped the course were excluded from this study. The course is divided into seven (7) modules and runs for 10 weeks.

The students’ access patterns were assessed based on the following variables:

1. Number of students who accessed (Views and Posts) to Self-Introduction Forum
2. Number of students who accessed (Views and Posts) to Announcement Forum
3. Number of students who accessed (Views and Posts) Discussion/Collaborative Forums
4. Number of students who accessed Course materials (Course Guide, Study Guide, Course Modules)
5. Number of students who submitted/completed requirements (assignments/exam/quiz/project)

The student performance variable was based on the final grade of the student. Student performance is categorized into four categories based on their final grades: High Performing Group (Final Grade: 86%-100%), Average Performing Group (Final Grade: 72%-85%), Low Performing Group (Final Grade:71%-60%), and Failed (Final Grade: less than 60%).

Data Collection

This study adapted the process in data mining by Romero, et al. (2007) (see Fig.1) as a tool for data collection and analysis. Log records of students’ activities were extracted from the LMS to assess the students’ access patterns and performance.

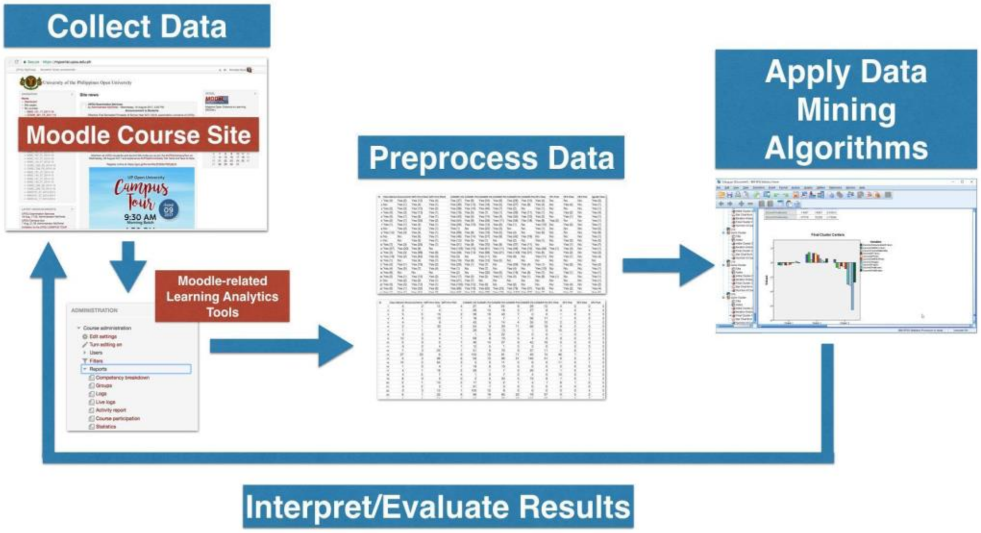

Collect Data
Moodle Course Site
Moodle-related Learning Analytics Tools
Preprocess Data
Apply Data Mining Algorithms
Interpret/Evaluate Results

Figure 1. Process of Data Mining in Moodle Course Site Using Standard Report Plugins in Moodle

1. **Collect data.**

This study adapted the process in data mining by Romero, et al. (2007) (see Fig.1) as a tool for data collection. Transcripts of messages and log records of students' activities including forum posts were extracted from nine asynchronous discussion forums in an online graduate course. This particular study uses the students' demographic information (Name, Sex, Location), and the data generated using a selected standard report plugin for Moodle which is available from the course site (see Table 2).

Table 3

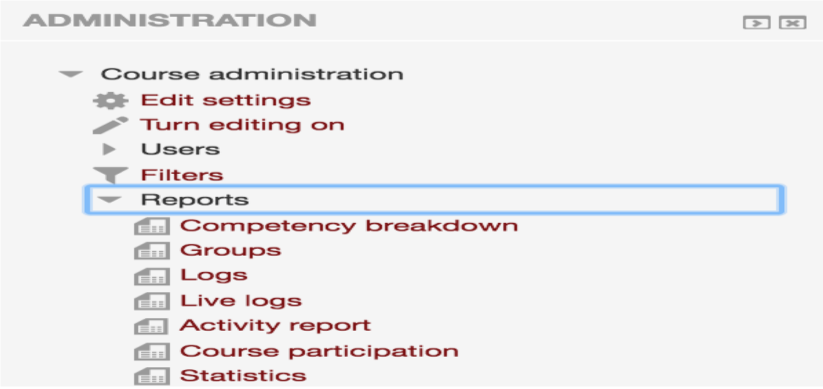
Selected standard report plugin for Moodle (Source: www.moodle.org)

Plugin	Useful for	Description
Logs	Teachers, Administrators, Decision-makers	Filterable log of events
Activity	Teachers	View count of activities in the course
Course Participation	Teachers	Single student's participation in the course
Statistics	Teachers, Admins	The statistics graphs and tables show how many **hits** there have been on various parts of the site during various time frames.

2. **Preprocess the data.**

The following data were generated using the selected standard report plugin for Moodle and were cleaned and preprocessed using Microsoft Excel (see Fig. 2). For this study, data for the following activities within the course site were generated:

1. Number of students who accessed (Views and Posts) to Self-Introduction Forum
2. Number of students who accessed (Views and Posts) to Announcement Forum
3. Number of students who accessed (Views and Posts) Discussion/Collaborative Forums
4. Number of students who accessed Course materials (Course Guide, Study Guide, Course Modules)
5. Number of students who submitted/completed requirements (assignments/exam/quiz/project)



ADMINISTRATION

- Course administration
  - Edit settings
  - Turn editing on
  - Users
  - Filters
  - Reports
    - Competency breakdown
    - Groups
    - Logs
    - Live logs
    - Activity report
    - Course participation
    - Statistics

Figure 2. Screenshot of the standard report plugin of the online course site

3. Apply data mining.

Preprocessed data gathered for this study were analyzed using SPSS. Descriptive statistics were used to determine the demographic characteristics and online participation behaviors of the students. Inferential statistics were used to investigate relationships among the variables.

Clustering is an unsupervised method for grouping, and it groups the data into sets of related observations or clusters (Myatt, 2007). This study used the SPSS to analyze and categorize the data using the K-means algorithm with a value of 4 to the number of clusters. K-means clustering approach was used to determine the usage patterns and clusters of online learning behaviors that emerge upon mining students' online participation data related to online learning activities. K-means clustering is an example of a non-hierarchical method of grouping a data set (Myratt, 2007). Computed values were visualized by using graphical and tabular representations.

4. Interpret, evaluate, and deploy the results.

The data in this study were interpreted and analyzed and conclusions were made based on the results.

RESULTS AND DISCUSSION

Demographic characteristics of students

To determine the online learners’ demographic characteristics and access to the course, data extracted were measured and analyzed using SPSS. Results revealed that out of 23 male students, more than half (60.87%) belong to the high and average performing groups while out of 22 female students, 77.27 percent belong to the high and average performing groups (see Table 4).

Table 4

Descriptive statistics and test of significance of demographic variables by performance

Performance	Frequency	Sex		Location	
		Male	Female	Overseas	Philippine-based
		23	22	6	39
High	14	26.09%	36.36%	33.33%	30.77%
Average	17	34.78%	40.91%	50.00%	35.90%
Low	9	26.09%	13.64%	16.67%	20.51%
Failed	5	13.04%	9.09%	0.00%	12.82%
p-value		0.677		0.688	

In addition, results also showed that all of the overseas students are all passers while 12.82% of Philippine-based students failed the course. A visual representation of the comparison of demographic variables by performance status (Figure 3) revealed that overall, most (78.43%) of the students passed the course while the rest (21.57%) did not.

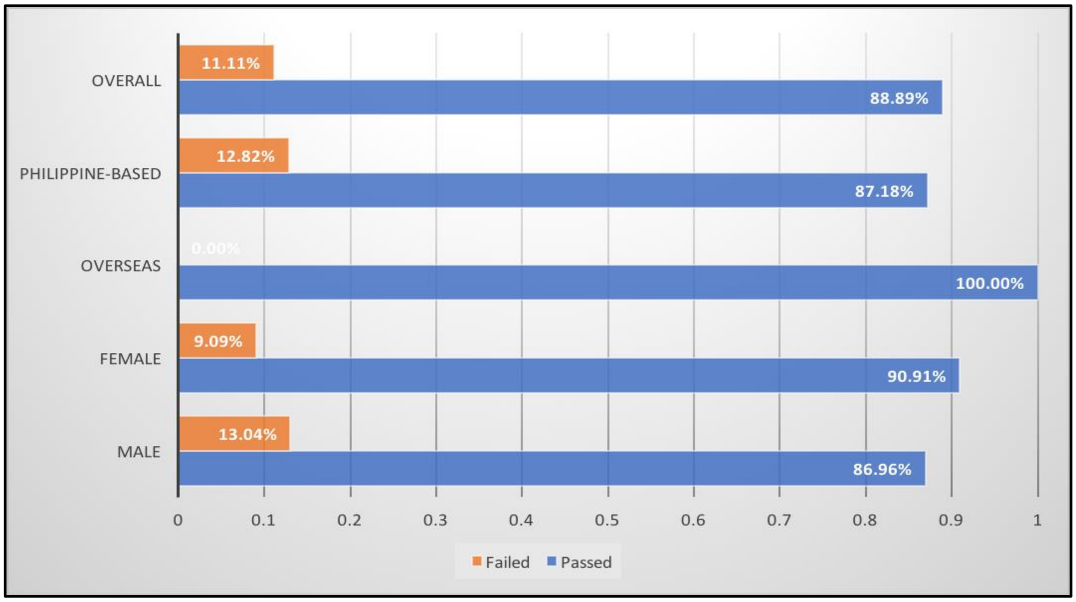


Figure 3. Comparison of demographic variables by performance status

The overall pattern of students' access to the course site

Descriptive statistics of students' access to the course site is shown in Table 5. Results revealed that out of the seven variables on students' access to the course site, the variable "Number of views to discussion and collaborative forums", got the highest mean value of 127.46, while the variable "Number of posts to self-introduction forum" got the lowest mean value of 2.28.

This suggests that students are more engaged if given more opportunities for collaboration. The variables "Access (views and posts) to self-introduction forum" got a total number of 1357 access (views and posts). This suggests that students have a high interest in knowing information about their co-learners. The variable "No. of course submitted" has the smallest standard deviation value of 1.04.

Table 5

Descriptive statistics of student's access to the course site

Variables	Total	Min	Max	Mean	SD
No. of views to announcements	237	0	27	5.26	6.35
No. of views to self-introduction forum	1254	0	159	27.86	40.38
No. of access to course materials	433	1	39	9.60	7.86
No. of views to discussion and collaborative forums	5736	10	768	127.46	135.84
No. of posts to discussion and collaborative forums	1064	0	78	23.64	17.59
No. of posts to self-introduction forum	103	0	11	2.28	2.63
No. of course requirements submitted	241	1	6	5.35	1.04

Figure 4 shows a visual presentation of the total number of students' access to the course site. results showed that the variable "Number of views to discussion and collaborative forums", got the highest total number of access/views of 5,736 (63%), while the variable "Number of posts to self-introduction forum " got the lowest total number of views of 103 (1%).

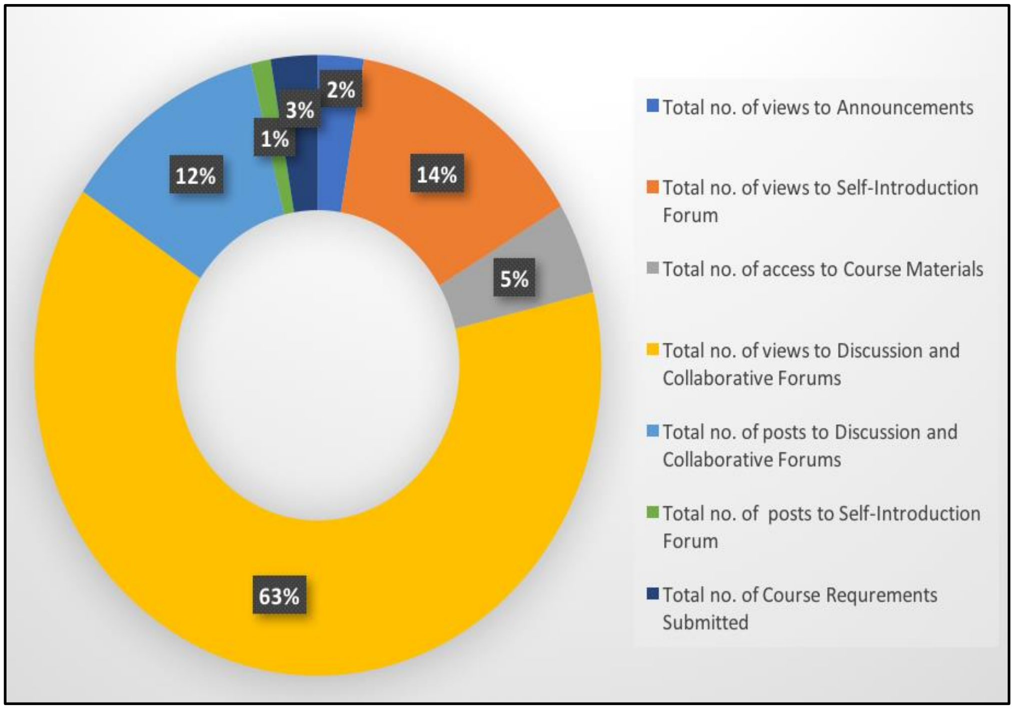


Figure 4. Students' access to the course

Comparison of students' course participation by week

To determine the course participation behavior of the students on a 10-week course, data logs were generated from the LMS. A visual representation of the course participation behavior per week is illustrated in Figure 5. Based on the results, there is a significant increase in the number of views and posts on Week 2 (14.56%), Week 4 (11.71%), Week 7-9 (33.4%), and Week 10 (17.71%).

Week 2 is the week where Assignment#1 has been posted. results also showed that this is the week with the second-highest number of access/views (15.42%), with Week 10 being the highest (20.02). Week 4 is the week when taking Quiz#1 is required. Week 7-9 is the period of submission of their Final Project. These periods are also the highest number of posts (49.10%). Week 10 is their Final Exam Week.

Results showed that this period has the highest number of views (20.02) which suggests that students are more active in accessing/viewing the course site during this period, probably to review course materials, discussion forums, and activities in preparation for their Final Exam. This suggests that students are more active when there are major activities/requirements in the course.

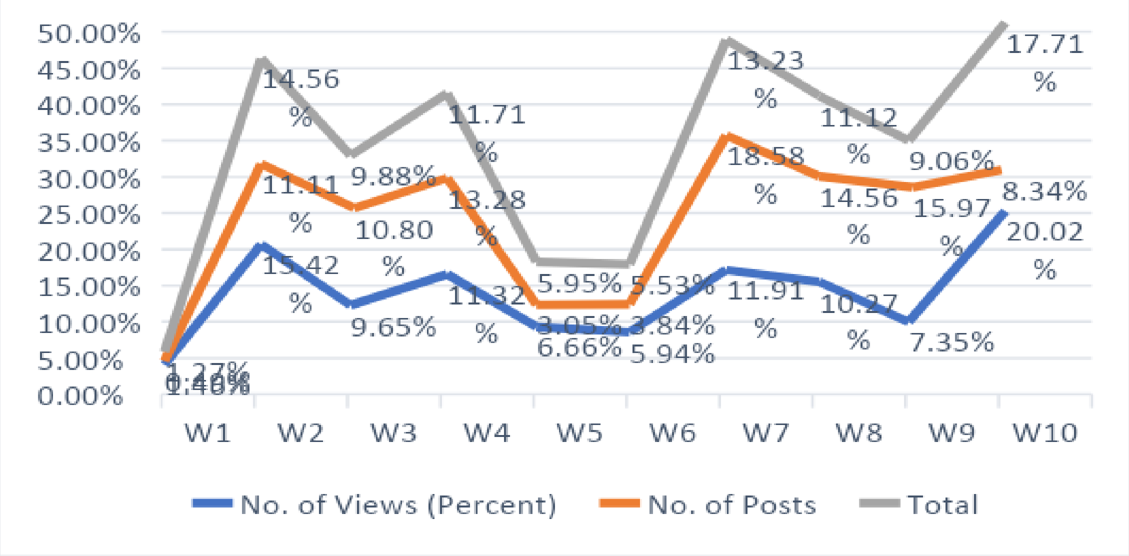


Figure 5. Comparison of Course Participation by Week

Clustering

Clustering techniques were applied to assess students based on their online participation characteristics. Clustering is an unsupervised method for grouping and it groups the data into sets of related observations or clusters (Myatt, 2007). This study utilized the following variables to describe and classify the characteristics of the students:

1. Number of students who accessed (Views and Posts) to Self-Introduction Forum
2. Number of students who accessed (Views and Posts) Announcement Forum
3. Number of students who accessed (Views and Posts) Discussion/Collaborative Forums
4. Number of students who accessed Course materials (Course Guide, Study Guide, Course Modules)
5. Number of students who submitted/completed requirements (assignments/exam/quiz/project)
6. Final Grades

Table 6 shows the demographic characteristics of students in four clusters.

Table 6

Demographic characteristics of students in four clusters

Variables		Cluster			
		1 N=2	2 N=20	3 N=15	4 N=8
Performance	High	7.14%	64.29%	0.00%	28.57%
	Average	5.88%	58.82%	17.65%	17.65%
	Low	0.00%	11.11%	77.78%	11.11%
	Failed	0.00%	0.00%	100.00%	0.00%
Sex	Male	4.35%	47.83%	34.78%	13.04%
	Female	4.55%	40.91%	31.82%	22.73%
Location	Overseas	33.33%	16.67%	16.67%	33.33%
	PH-Based	0.00%	48.72%	35.90%	15.38%

Table 7

Means for clustering results

Variables	Cluster			
	1	2	3	4
No. of views to announcements	21	5.1	3.4	5.25
No. of views to self-introduction forum	147	16.45	9.33	61.25
No. of access to course materials	22.5	7.2	7.6	16.25
No. of views to discussion and collaborative forums	610.5	11.5	55.8	172.13
No. of posts to discussion and collaborative forums	44.5	23.25	12.27	40.75
No. of posts to self-introduction forum	6.5	1.25	1	6.25
No. of course requirements submitted	6	5.9	4.27	5.88
Final Grade	87.16	82.99	60.1	83.11

As shown in Table 7, the students were classified into 4 clusters with 2, 20, 15, and 8 students, respectively. Cluster 1 is characterized by students with the highest level of access to the course site and has an average final grade of 87.16. Cluster 2 is characterized by students with a low mean value in accessing the course materials (7.2) and a number of views to discussion and collaborative forums (11.5), with an average final grade of 82.99. Cluster 3 is characterized by students with the lowest level of access to the course site and has the lowest average final grade of 60.1. Cluster 4 is characterized by students with an average level of access to the course site and an average final grade of 83.11. A visual representation of the means for clustering can be seen in Figure 6.

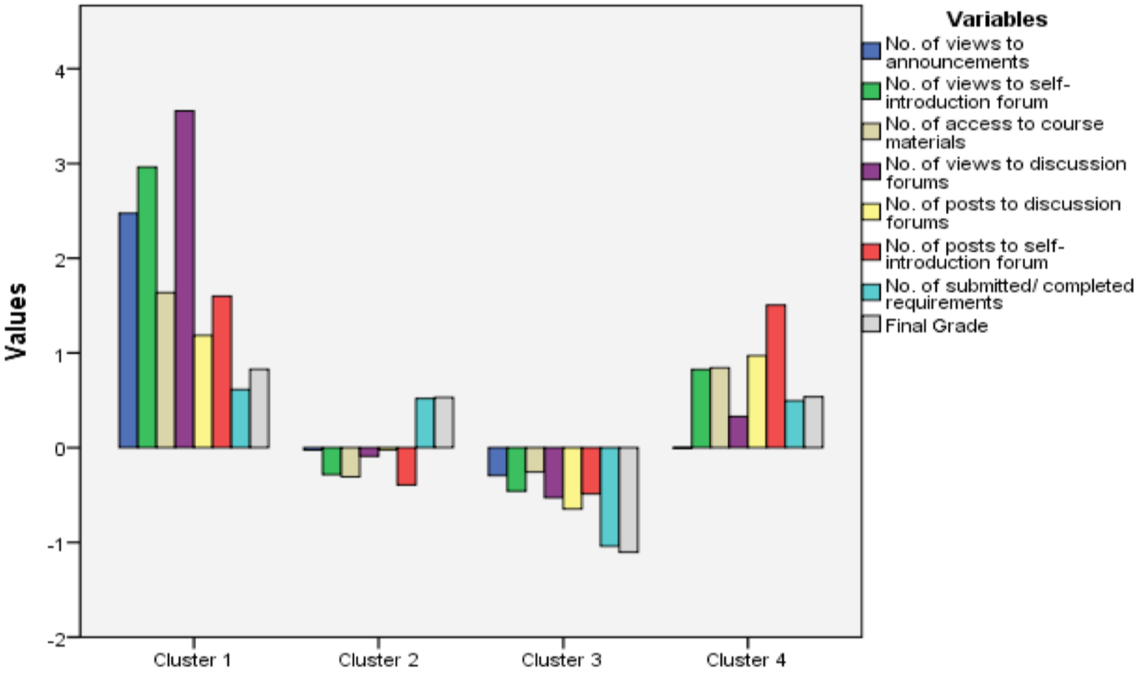


Figure 6. Visualization of means for clustering results

CONCLUSION

There is a huge amount of data that an LMS can capture and store, making it difficult for teachers to monitor. This study showed the advances in employing the combination of cluster analysis and data mining in assessing the students' online access patterns and performance. With data mining techniques, access patterns of students based on demographic variables such as gender and location were identified in relation to the students' performance. It is interesting to know that in this particular research study, female students performed better than male students in terms of their final grades.

The overall access pattern of the students suggests that students are more engaged if given more activities and opportunities for collaboration such as in discussion forums. Results show that a big part of students' access to the course site is to access (view and post) the discussion forums and view the introduction forum where they learn something from their classmates.

Results also show that there is a high level of access to the self-introduction forum, suggesting that students are interested to know information about their co-learners. This suggests that the teacher in charge can add more activities that facilitate more interesting discussions about themselves among students. Results also showed that students are most active in accessing the course site on the second week of classes, on the week of the Final Exam, and on those that require them to submit or complete assignments.

The clustering approach, as applied in this study, helped in classifying different groups of students according to their characteristics which can be the basis for course improvement of the teacher in charge. This study suggests that the data mining techniques, particularly the statistics, clustering approach, and visualization can be applied to analyze the data generated using the selected standard report plugins for Moodle which are available from the course site.

This study recommends further studies be done in other courses to check the consistency of the result. It is also recommended to conduct a survey of the students to determine their study habits and their perception of the effect of their location on their performance in class. Further study in the subject matter relating to other learning behavior of students in an online class is highly recommended. The use of other learning analytics such as reports, blocks, and plugins to assess the performance and learning behavior of students is recommended as well.

Further study on the use of multi-analysis methodologies to investigate the interaction in terms of posts accessed and viewed in the discussion forum is also recommended by the researcher.

REFERENCES

Alfiani, A. P., & Wulandari, F. A. (2015). Mapping Student's Performance Based on Data Mining Approach (A Case Study). *Agriculture and Agricultural Science Procedia*, *3*, 173-177.

Bagarinao, Ricardo T. (2011). Learners' Access Patterns and Performance in an Online Course in Science, Technology and Society. ASEAN Journal of Open Distance Learning. Retrieved from http://ajodl.oum.edu.my/sites/default/files/document/vol3-no1/Vol3-01.pdf.

Butrous, N. (2011). Online access patterns and students' performance. *Journal of Systemics, Cybernetics, and Informatics, 9(2),* 74-78. Retrieved 20 July, 2017 from http://www.iiisci.org/journal/CV$/sci/pdfs/OL268TF.pdf

Dimopoulos, Ioannis and Petropoulou, Ourania and Boloudakis, Michail and Retalis, Symeon (2013) *Using Learning Analytics in Moodle for assessing students' performance.* In: 2nd Moodle Research Conference (MRC2013), 4th and 5th October, 2013, Sousse, Tunisia.

Haig, T., Falkner, K., & Falkner, N. (2013, January). Visualization of learning management system usage for detecting student behaviour patterns. In *Proceedings of the Fifteenth Australasian Computing Education Conference-Volume 136* (pp. 107-115). Australian Computer Society, Inc..

Hung, J. L., & Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. *MERLOT Journal of Online Learning and Teaching*.

Mohamad, S. K., & Tasir, Z. (2013). Educational data mining: A review. *Procedia-Social and Behavioral Sciences*, *97*, 320-324.

Myatt, G. J. (2007). *Making sense of data: a practical guide to exploratory data analysis and data mining*. John Wiley & Sons.

Park, Y., Yu, J. H., & Jo, I. H. (2016). Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute. *The Internet and Higher Education*, *29*, 1-11.

Romero, C., Ventura, S., & García, E. (2007). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368-384.

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, *42*(1), 85-106.