Analysis of Student Learning Behavior via Moodle Data

Narine Manukyan
ASNET-AM
Yerevan, Armenia
e-mail: nana@sci.am

Parandzem Hakobyan
IIAP NAS RA
Yerevan, Armenia
e-mail: par_h@iiap.sci.am

Arusyak Harutyunyan
ISEC NAS RA
Yerevan, Armenia
e-mail: arusyak.harutyunyan@isec.am

***Abstract*—Moodle LMS has a variety of statistical and data analysis toolkits for management and monitoring of student progress and behavior, whereas their consistent use by faculty and administrators shall help improve the learning process. Nevertheless, simply counting and visualizing the numbers of posts and clicks are, as a matter of fact, useless for learning process management and its outcomes. There are certain tasks and statements of questions, which require data analysis with the application of more efficient software. The article examines joint online MBA log data via Moodle server on online.isec.am. For over 5 years, the online platform has accumulated a bulk of data from log files obtained in the course of this program, which detail the activities of learners. The aim of this article is to analyze learning activities and learning behavior of online MBA students with the application of R software, which can help to perform thorough analysis of the data by applying statistical techniques, data mining and data visualization.**

***Keywords*—Distance learning, learning management system, Moodle, learning analytics, statistical method, correlation, Kruskal-Wallis test.**

I. Introduction

Human activity can no longer be imagined without large-scale data collection and analysis. Diverse and comprehensive data collection has occupied its unique place in education, educational data processing, and analysis have turned into an integral part of the amelioration and management of educational processes. The term "datafication" was first used by Mayer-Schönberger, V. and Cukier, K. in 2013, which describes the transformation of non-structured human data into a structural form, which is suitable for processing and analysis in various information systems [1].

Learning management systems (LMS) comprise four main components: people, data processes and information technologies.

The accumulation of data in information systems and their further processing enables system participants to improve further processes and ensure the stable operation of the system. Higher education institutions, especially during the epidemic, using their information systems and VLE platforms, willingly or unwillingly accumulated a large amount of data, whereas their analysis opens up great opportunities for the entire higher education system.

The purpose of this study is to analyze the log files of Moodle Virtual Learning Environment for Joint Online MBA Program administered by International Scientific-Educational Center of NAS RA and Public Administration Academy of RA to determine the degree of student participation in various course components for the further modernization of distance/online learning process [2].

Within the framework of "Business Administration (Management)" online program under "Management" academic program, 26 subjects were developed and launched, each having a duration of 16 weeks. Lecturers and coordinators made use of over 80% of the interactive components and resources in the Moodle system to design and create distance-learning/online resources. The courses are designed with a high level of autonomy, which, in turn, provides an individualized approach for each student. The educational program of "Masters of Business Administration" consists of five subject blocks.

II. Data and Methods

This paper examines Moodle LMS event logs for "Digital Marketing" course taught in the second semester of 2017-2021 academic years. All the activities and events are recorded in the logs. Event logs are stored in databases and allow information to be collected by any participant in the system: student, lecturer, administrator, coordinator, etc.
The collected data contains information about the activities when visiting the course and activities related to the course components in the given time period. Based on these logs, the Moodle system generates various reports. The event log contains the following information:

- Time - the time activity is carried out,
- Username Name - surname of the user carrying out the activity,
- Affected user - the name of the related user if the activity concerns another user,
- Event context - the course where the activity was carried out,
- Component - the course in which the activity was carried out,

- Event name - the type of activity (e.g., view, delete, create, update),
- Description - detail information regarding the activity,
- Source - the source of the activity,
- IP Address - the network address (IP) of the device on which the activity was carried out.

Only the time of the event is recorded in the logs, nevertheless, its duration is not mentioned, therefore it is impossible to determine exactly how much time the student spends studying this or that material, how much time the student works with the resource can only be guessed when s/he goes to the next activity. Event logs can also show which resources have the high demand of the students and which ones have never been accessed.

In this particular case, Moodle LMS event log comprised data collected for 3 groups of students (a total number of 35 students) collected for 2017-2019, 2018-2020 and 2019-2021 academic years.

Predicting student progress based on the data collected is difficult because an online master's program has classroom components, so predictions are only available after the course is over [3].

To determine the degree of students' involvement in the different components of the course with the overall progress of the subject, we have selected 3 compulsory components of the examined courses from the logs – those of 8 BigBlueButton webinars, 16 assignments and participation in 16 forums.

The non-parametric methods have been used during statistical analysis because normal distribution conditions had not been satisfied for the data provided [4]. R language has been used for the data manipulation, statistical analysis, calculations and graphical representations [5].

III. Analysis and Discussion

In order to perform statistical analysis, it is necessary to filter the data needed for the study from the obtained data and present them in a single data table. In other words, pre-processing the data requires a lot of time and a number of actions, for which a modern *dplyr* package has been considered in parallel with a number of functions of the *base* package.

Thus, the process of data conversion and processing consists of the following steps:

- Removing data not studied within the given analysis and related only to system component;
- Removing duplicate entries;
- Removing activities related to lecturers, administrators and tutors;
- Separating the data on the participation of the students in the webinars, forums and completed assignments according to the years. The description of the columns in the final table is given in Table 1.

Appropriate correlation coefficients were calculated in order to understand to what extent participation in webinars, assignments, and forums contributes to a student's progress/grade.

Table 1. Lists of Attributes after Transformation.

Column Names	Description
Username	Name of the Student
Course	Names of course (in accordance with academic years)
N_BBB	Number of participations in webinars
N_ASS	Number of completed assignments
N_Forum	Number of forum posts
Grades	Final course grade

The correlation matrix plot /heatmap/ of the attributes of the webinars, assignments and forums with final grades helps to identify which of the attributes are correlated so that we can understand the influence of selected factors on the final grade.

The Spearman's rank correlation coefficient was calculated because these data are not normally distributed according to Shapiro-Wilk test, so the calculation method of Pearson's correlation coefficient is not applicable.

Fig.1a) - Fig.1d) show the diagrams of the scattered quantities of the grades and relevant webinar and forum participation and assignment completion. Visually, we can assume that there is an interconnection, but we can have confidence in the existence of the interconnection only if the values of the correlation coefficients and their significance are substantiated.

Fig1d) presents a correlation heatmap, where we can see that there is a strong correlation between student progress and participation in relevant webinars, assignments and participation in the forums. This means that the increase in participation in webinars and forums, as well as the completion of assignments leads to an increase in the students' grades. Testing the significance of these correlation factors substantiated that they were significant (p-value<0.001). The coefficients of pair correlation among the participation of webinars, the completion of assignments and the participation in the forums are visible in the correlation heatmap, which are also significant. In other words, there is a positive interaction between these attributes. The strongest correlation is between grades and assignment completion. The coefficients of partial correlation between these two quantities were calculated according to the participation in the forums and webinars, being equal to 0.38, which is smaller than the actual correlation coefficient - 0.78. This suggests that the factor of participation in webinars and forums strengthens the relation between grades and assignments.

The next task is to find out whether the progress of the students is, on average, the same or there is a change observed during these three academic years. In other words, whether the factor of years of teaching the subject exerts effect on progress.

Fig. 2, shows the boxplots of the observed attributes by years. Here, the horizontal line in each of the boxes corresponds to the data median, whereas the black colored dots are the mean values of these data.

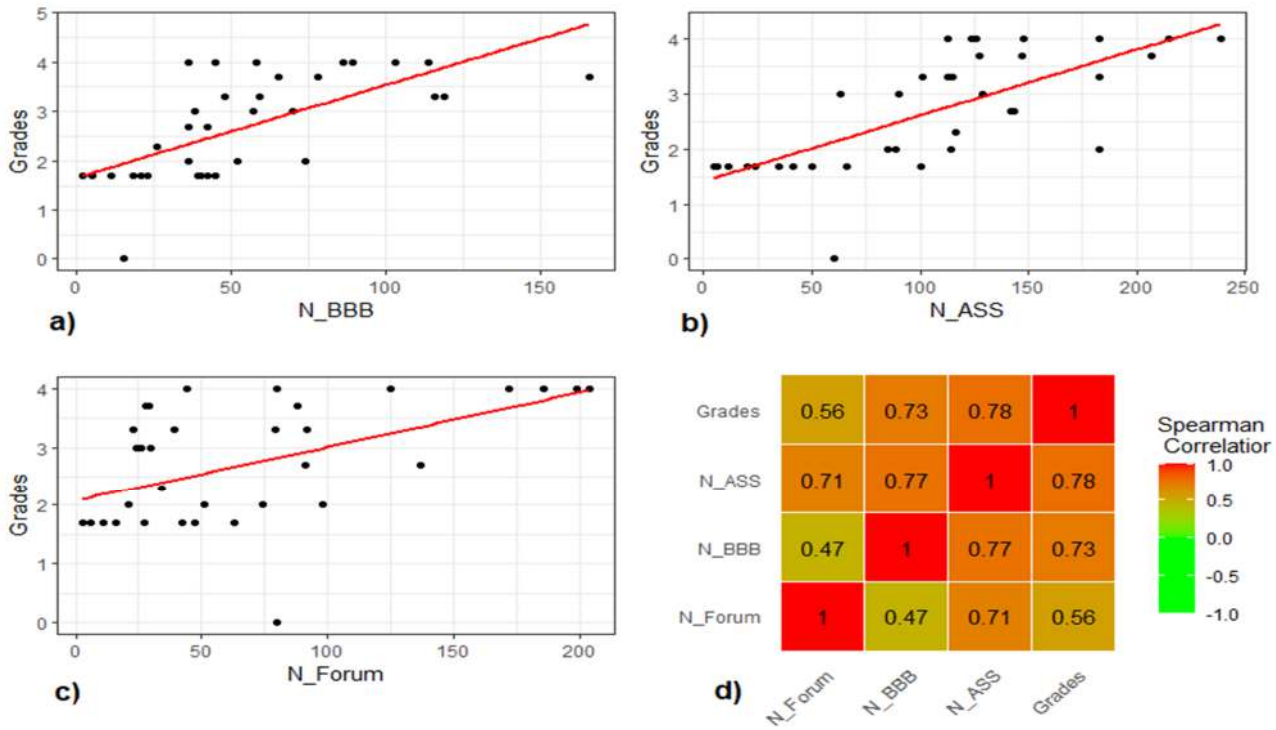

Fig 1. Scatter plot and correlation heatmap of considered quantitative attributes.

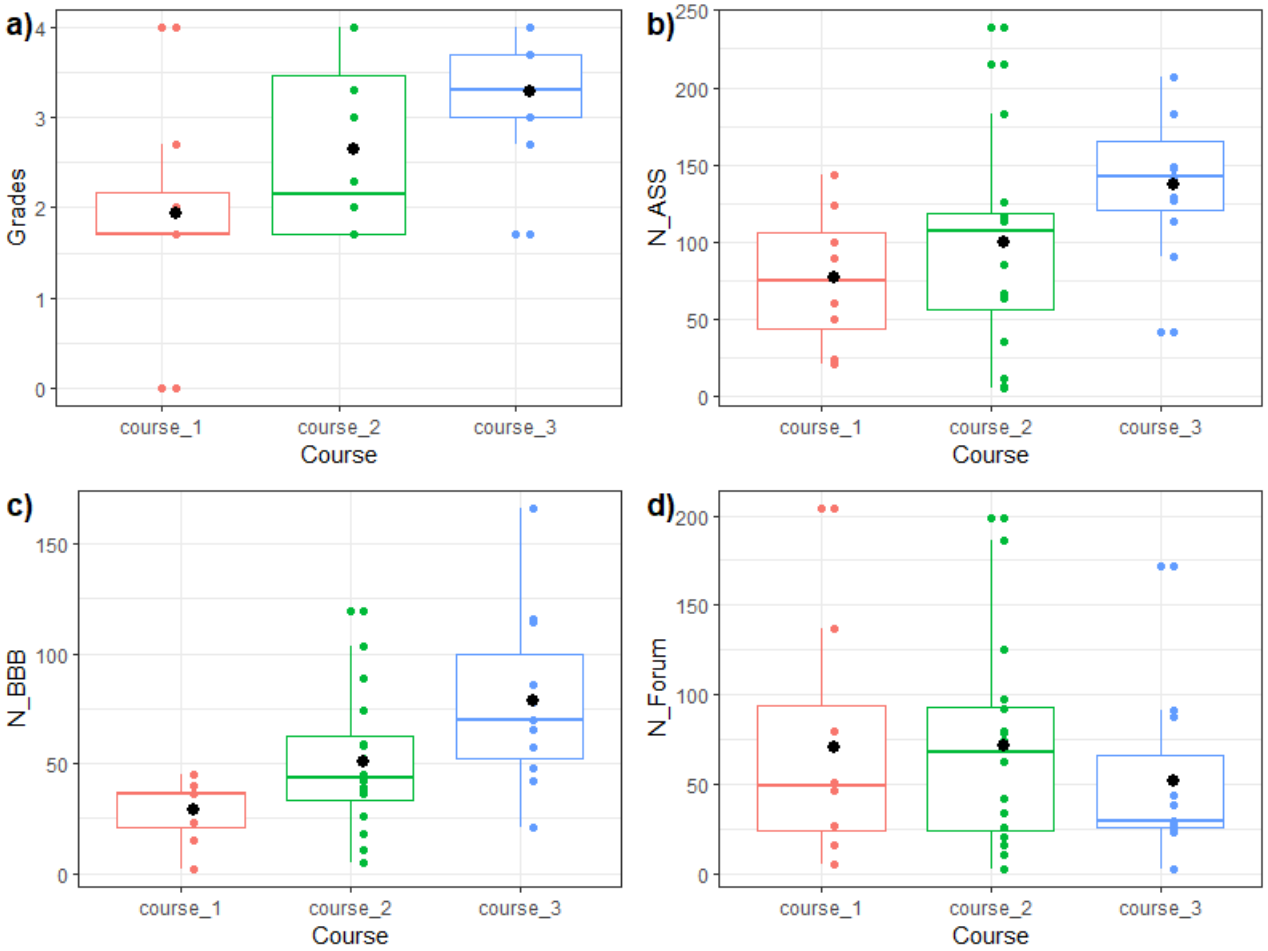

Fig 2. Boxplot of considered quantitative attributes by years.

Judging by visual Figure 2a we can say that we have an average progress over the years constituting 1.94, 2.65, 3.28. Now we should make sure that the difference is statistically significant.

One-way analysis of variance (One-way ANOVA) is not considered as one of the parametric methods for this case compared to the mean, as the conditions for its application are not met with respect to the normal distribution of group data and the variance equation. Therefore, its non-parametric analogue Kruskal-Wallis test was considered [6]. As a result of the testing, the significance of the difference between the mean was substantiated (p-value = 0.03642).

To find out among which specific groups there are statistically significant differences observed, and among which groups there are not, the Dunn's test of multiple comparisons was observed [7]. Under this test, a statistically significant difference by significance level of 0.05 is observed only in the progress of 2017-2018 and 2019-2021 academic years. In other words, we have positive dynamics in terms of student progress.

Graphs b) and c) in Figure 2 show an increase over the years in terms of the number of students participating in webinars and completing assignments. As in the above version, the analysis revealed that there is no significant difference in terms of the number of assignments, whereas there is a difference observed between 2017-2018 and 2019-2021 academic years in the case of participation in webinars.

Figure 2d) shows that the participation in the forums in 2019-2021 is lower than the visual average compared to the others, but it is not significant, which was substantiated by calculations using the methods observed in the previous cases. Thus, there is an increase in progress in terms of the participation in webinars, nevertheless, there is no change observed in terms of participation in the forums and completion of assignments.

IV. Conclusion

The comprehensive analysis of data on the behavior of students and lecturers in the system participants of the study showed that there has been an increase observed in student progress over the years. This may also be possible due to the use of new teaching methods. It needs additional research in order to obtain the relevant data.

Interest in webinars is increasing, but there is no significant change in assignment completion, as well as students have low engagement in forums.

However, students' progress is strongly positively correlated with their participation in the webinars and assignments, which show that they are factors in promoting students' progress.

References

[1] V. Mayer-Schönberger and K. Cukier, *Big data: a revolution that will transform how we live, work, and think.* Boston: Houghton Mifflin Harcourt, 2013.

[2] A. Harutyunyan and N. Manukyan, "First initiative of higher education digitization in armenia: issues and challenges", *Proceedings of the CSIT Conference.*

[3] R. Joseph Quinn and Dr. Geraldine Gray, "Prediction of student academic performance using Moodle data from a Further Education setting", *Irish Journal of Technology Enhanced Learning*, vol 5, no. 1, 2020.

[4] P. Sprent and N. C. Smeeton, *Applied Nonparametric Statistical Methods*, Chapman and Hall/CRC, 3rd ed., 2020.

[5] R. Kobaccof, *R in Action, Data Analysis and Graphics with R*, Manning Publications Co. 20 Baldwin Road PO Box 261 Shelter Island, NY, 2rd ed., 2015.

[6] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis", *Journal of the American Statistical Association*, vol. 47, pp. 583–621, 1952.

[7] A. Dinno, "Nonparametric pairwise multiple comparisons in independent groups using Dunn's test", *The Stata Journal*, vol. 15, no. 1, pp. 292–300, 2015.