Cluster analysis of students' activities from logs and their success in self-assessment tests

Danijel Filipović, Igor Balaban, Dijana Oreški
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
{dfilipovi, ibalaban, dijoresk}@foi.hr

Abstract*. Many educational institutions, especially the ones in higher education, implement Learning Management System (LMS) to assist teachers in teaching and aiding the students in learning. Some of such features available to teachers include logging student's activity and creating custom tests. This paper aims to: (1) Identify relationship between the student's activity on the course and their success on the self-assessment tests; and to (2) Create behavioural profiles of students. In order to do so, the correlation analysis and cluster analysis are performed on a data set retrieved from a course implemented on the Moodle LMS at the University of Zagreb, Faculty of organization and informatics. The research results indicated the relationship between students´ activities and their performance, as well as the profiles of students based on their activities and success on tests.*

Keywords. Cluster analysis, student activity, student success, logs, LMS

1 Introduction

The LMS allows teachers to customize their online or blended courses according to their needs. Those systems are able to record every student's activity, as well as their progress on the course (e.g. their success on tests). "When accessing an LMS with their personal account, students create a digital profile that is saved in the LMS log files" (Kadoić and Oreški, 2018). A log is a list of students' events where a single line represents a timestamp and fields give information about the activity performed (Romero et. al, 2013). These data represent a valuable source for various research activities. In order to gain an insight into such data, the data analytics is performed.

The research is based on the data derived from the blended course "Business Informatics". This course is being conducted at the University of Zagreb, Faculty of Organization and Informatics, at the undergraduate vocational study programme PITUP - Information Technology in Business Application. Business Informatics is taught within four study centres in Croatia: PITUP Varaždin, PITUP Križevci, PITUP Sisak and PITUP Zabok. The data is from the academic year 2017/2018.

The aforementioned course contains several self-assessment tests that students can take. The tests are optional and can be retaken. These tests merely serve as tools that students can use to assess their current knowledge on the subject as a complement to their learning or to improve possible gaps in their knowledge.

We will combine the activity data from students' log files and their success on the self-assessment tests to identify a correlation between students´ activities and their results. Furthermore, we will apply cluster analysis on the data set in order to explore profiles of students' behaviour.

The paper is organized as follows. In the second section we provide a brief review of the previous research on the given topic. Section 3 explains the methodology and section 4 describes the data used in the research. In the Section 5 we present research results and interpretation. We conclude in Section 6.

2 Related work

The LMS provides opportunities for an enhancement of student learning and, consequently, can impact students' final grades (Pislaru and Mishra, 2009). Various authors investigated a relationship between students' LMS activities and their performance.

Asif et al. (2017) used clustering to analyse the students' academic progression in a 4-year bachelor's degree programme. Students were assigned into three clusters based on their average marks: *Low*, *Intermediate* and *High*. They also used decision trees to predict the performance of their students at the end of the semester. Their likely performance, or their final grade, is predicted based on the grades achieved in the first two years.

Alfan and Othman (2005) analysed the performance of students from the University of

Malaysia that took the courses in business and accounting programme. In order to analyse students` performance, the authors performed various analyses. Their aim was to answer several research questions: does previous knowledge of the subjects taught at the study programmes affect their final CGPA, is there a difference in performance between male and female students and whether the students' performance is dependent on their race – Chinese, Indian and Malay?

Bouchet et al. (2013) analysed students' interaction with MetaTutor, the multi-agent intelligent tutoring system. Students were randomly assigned into two test groups - *Prompt and Feedback,* where the students were prompted by the MetaTutor to use specific strategies and were immediately given feedback, and *Control*, where students did not receive any prompts or feedback. The interaction data MetaTutor stored was extracted and clustered by applying Expectation-Maximization (EM) algorithm in Weka 3.

Saarela and Kärkkäinen (2015) analysed the performance of students at Department of Mathematical Information Technology (DMIT) at the University of Jyväskylä in Finland. They specifically focused on performance in Computer Science curriculum. First, they performed a correlation analysis to see if students' grades in certain courses affect their overall success. Furthermore, they performed a cluster analysis and analysed the clusters based on the average grade per course and the average credit score per cluster. Finally, they performed predictive analysis to infer which courses have the highest influence on students' performance.

Talavera and Gaudioso (2004) used clustering to obtain several behavioural profiles of students based on their log files recorded by the LMS. They applied the EM algorithm to dataset which generated 6 different clusters and, to an extent, 6 different behavioural profiles of students.

Gašević, Dawson, Rogers, and Gasevic (2016) explored the extent to which students´ activity influence the prediction of academic success in a blended learning model. LMS data included the usage of the Moodle features: forums, course logins, resources, assignments, book, quizzes, feedback, lessons and chat.

Wang, Lv, Cao and Biao (2017) collected the log data generated by students in the self-learning platform named "Engineering Mechanics Experiment" Autonomous Learning Platform which was designed by their own institution. In their paper, they analysed two set of factors: factors influencing students' landing behaviour and factors influencing students' resource browsing behaviour.

Cantabella et al. (2018) conducted a case study at Catholic University of Murcia in which they analysed the student behaviour in three different modalities (online, on-campus and blended) in the following academic years: 2012/2013, 2013/2014, 2014/2015 and 2015/2016. The analysis was performed with the help of a framework that is built with big data technologies – Apache Hive for storing the student data and Apache Hadoop for performing various statistical analyses. The data was collected from the Sakai LMS, specifically the events triggered by students where extracted and stored in Apache Hive. First, they ranked the tools the students used in Sakai LMS, specifically they measured how much each tool was used for each academic year for each modality. Second, they ranked the total amount of events the students triggered for each modality. Third, they analysed the relationships between the events for each modality. In this instance, the relationship indicates the combination of events students triggered in the same session (Cantabella et al., 2018, page 22). Lastly, they searched for monthly and weekly connection trends - the number of times students visited Sakai LMS - across all years for each modality.

Estacio and Raga Jr. (2017) tried to show whether students' learning behaviour can be extracted from logs recorded by Moodle LMS and visualized accordingly. They also tried to determine if the aforementioned logs can give insight into students' course performance and if their demographic profile affects their level of activity on Moodle LMS. The interesting part of their research is the application of Vector Space Model algorithm to extraction and visualization of students' learning behaviour.

Based on the results of previous research, we have defined the following research goals and research questions. The aim of this research is three-fold:

(i) to identify a relationship between the students' activities within the course and their success on the self-assessment tests
(ii) to create behavioural profiles of students that took the blended course
(iii) to investigate the differences between students from different study centres.

The following research questions were set up:

RQ1: Is there a correlation between students' activities and their success on self-assessment tests?

RQ2: What are the profiles of students' behaviour on the course?

3 Methodology

In order to answer the first research question (RQ1), the correlation analysis was performed on the full dataset, which included students from all PITUP centres. The correlation analysis answered which attributes are connected to the student performance.

With the aim to answer the second research question (RQ2) we applied the cluster analysis. Clustering is a process which groups objects into classes, or groups, of similar objects (Romera & Ventura, 2007). Clustering is a type of unsupervised learning algorithm. According to Baker (2010), clustering is one of five general methods that can be used in educational data mining (EDM). The literature review revealed the applicability of cluster analysis in analysing students' behaviour and progression. Bovo et. al. (2013) proved that cluster analysis is a great tool for profiling students based on LMS data.

Based on the log data of students' activities from all PITUP centres, we separated the data into four datasets (one for each PITUP centre) and applied the clustering for each one.

Open-source data mining software Weka 3 was used for clustering. As for the algorithm used in the clustering, a simple *k*-Means algorithm was chosen.

4 Data description

In this section, we explain the data used in the research, which was collected from Moodle course log in an actual class.

4.1 Log data

The students' log data was retrieved from the Moodle LMS implementation of the blended course Business Informatics. The original dataset, which was used to extract a new dataset for analysis, contains the attributes described in Table 1.

The log data contains activities from 356 students, of which 33 belong to PITUP Križevci, 74 belong to PITUP Sisak, 124 belong to PITUP Zabok and 125 belong to PITUP Varaždin. In total, this dataset contains 667 174 instances of log data. The earliest instance was created on October 2017, and the latest instance on May 2018.

Table 1. Attributes of the original log dataset

Name	Description
Time	Date and time the instance of log data was created
Full name	First and last name of the user that triggered the event that created the new instance of log data. The "user" can be either student, teacher or the system. In case of a system, a dash symbol (-) is used
Affects user	First and last name of the user on whom the specific event that triggered the event that created the new instance of log data affects. For example, if user *John Doe* views the profile of user *Jane Doe* this attribute will contain value *Jane Doe*. Otherwise, a dash symbol is used (-)
Context	Label of the specific Moodle page of the course that the user was viewing where the event that triggered the creation of the new instance of the log data occurred. It's usually the title of a lesson, title of a test, name of a file that was downloaded, etc.
Component	Category to which the *context* belongs, i.e. *Lesson*, *Test*, *File*, etc.
Name	Name of the event that triggered the creation of the new instance of log data
Description	Detailed description of the event that triggered the creation of the new instance of log data
Source	Source of the event. Only contains the value *web*
IP address	IP address of the user that triggered the event that created the new instance of the log data

Values from the attribute *Component* were used as a set of new attributes for the new dataset. The idea was to count the number of times each value of the attribute *Component* appears for each distinct student. This is where the attribute *Full name* is used – to group the total number of each *Component* by student's name. The instances of log data made by teachers and the system were omitted.

To actually group values of the attribute *Component* by students' names we used PivotTable functionality of Microsoft Excel. Thus, we got one part of the complete dataset – a student's activity dataset.

4.2 Success on self-assessment tests

The students' success on self-assessments tests data was also retrieved from the Moodle LMS implementation of the blended course. The system allows teachers to generate reports for any valid context: attendance, project grades, exam grades, success from online tests, etc.

For this research, we generated a report on each student's success on every self-assessment test, of which there are nine, and their average success on all of the tests combined. Thus, we got a second part of the complete dataset – a students' success on self-assessment tests dataset – whose attributes are explained in Table 2.

Table 2. Attributes of the students' success on self-assessment tests dataset

Name	Description
Student	Contains student's first and last name
SA 1-9	Contains student's success for each individual self-assessment test. Each of the self-assessment test is one attribute in the dataset (hence the 1-9 in the name).

	Contains nominal values ranging from 0,0 to 1,0
SA AVG	Contains student's overall success in all of the self-assessment tests. Contains nominal value ranging from 0,0 to 1,0

4.3 Final dataset

As explained, the dataset derived from the log data, as explained in section 4.1, and the dataset explained in section 4.2 are two parts of a complete dataset that was used in this research. To analyse the data, the datasets where converted into CSV files where the values are separated by semi-colons (;). Then, a simple Python script was written that merged the columns by student's full name and wrote it into a third CSV file. That CSV file is now the complete dataset used in the research.

However, due to the correlation analysis results explained in section 5, we removed the attributes *Selection*, *Folder*, *Records* and *SA 6* because they had no major effect on other attributes.

The attributes of the complete dataset are described in Table 3.

Table 3. Attributes of the complete dataset

Name	Description
Student	Contains student's first and last name.
File	No. of times the student downloaded a file or viewed it in-browser.
Forum	No. times the student viewed the forums section of the course.
Student report	No. times the student viewed his/her or other student's Moodle LMS profile.
Lesson	No. of times the student viewed lessons on the course.
File upload	No. times the student uploaded a file.
Link	No. of times the student clicked on an outgoing hyperlink.
Overview report	No. of times the student viewed their overview report in the gradebook.
Page	No. of times the student viewed any of the course's pages
System	No. of other general user activities
Test	No. of times viewed or took the online test
Homework	No. of times the student uploaded a homework file. In the Moodle LMS, this could relate to an actual homework (seminar papers, source code, etc.) or a file the student created as part of an exam
SA 1-5, 7-9	Contains student's performance for each individual self-assessment test. Each of the self-assessment test is one attribute in the dataset (hence the 1-9 in the name). Contains nominal values ranging from 0,0 to 1,0.
SA Average	Contains student's overall success in all of the self-assessment test. Contains nominal value ranging from 0,0 to 1,0.

4.4 Datasets per PITUP centre

For this research we wanted to perform the cluster analysis for each PITUP centre individually. That means separating the complete dataset into four smaller datasets which correspond to one of the PITUP centres.

To do this, we used the Moodle LMS to extract the list of students for each centre and created another Python script that takes the complete dataset and copies each row of data into one of four corresponding new datasets. After that, the data was prepared for the analysis.

5 Research results

The most important results from the log file analysis are presented in this section.

5.1 Correlation analysis

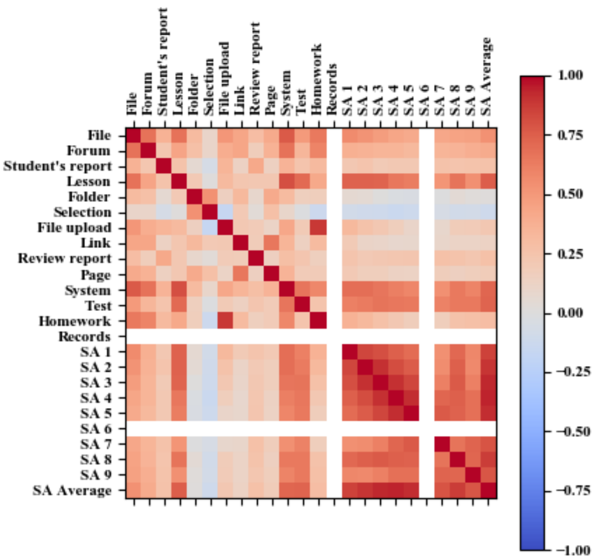

Figure 1. Correlation matrix

To determine whether there is a relationship between students' activities and their success on self-assessment tests, we computed a correlation matrix of the complete dataset. For the computation, another Python script was created, with the help of *Numpy*, *Pandas* and *Matplotlib* libraries, that generated a visualisation of the correlation matrix (Figure 1).

The correlation matrix shown in Figure 1 uses a heatmap to show direction and strength of the relationship and to display how much the two attributes correlate to each other. The redder the point (or square) of intersection is, the greater the correlation is. Opposite of that, the bluer the point of intersection is, the lower the intersection is.

From the correlation matrix we can see that attributes *Records* and *SA 6* seem to be excluded from the computation of the matrix. This is because for every instance in the dataset the values of those attributes were zero. Thus, we removed those attributes from the further analysis since they do not provide any information.

For other attributes regarding students' activities we can see that there are both weak and strong correlations with the attributes regarding students' success on the self-assessment tests. The attributes *Folder* and *Selection* have close to zero correlation with the success attributes, but due to their nature we determined that they can be ignored and removed from the complete dataset. On the other hand, attributes *Lesson*, *Test* and *System* have strong correlations with the attributes which indicate student's success.

High correlation of attributes *Lesson* and *Test* with the success attributes can be explained as follows:

- While studying, students use the self-assessment tests to test their knowledge. If they answered a question wrongly or didn't know the answer, they would recheck the lessons to see what the correct answer was and fill the gaps in their knowledge
- Attribute *Test* basically counts the number of times a student took the test. Since the tests can be repeated the number is higher and therefore the stronger the correlation with the success attributes.

We concluded that there is a correlation between students' activities and their success on the self-assessment test.

5.2 Cluster analysis

In order to answer the second research question, we have performed cluster analysis. Each dataset that corresponds to its PITUP centre was imported into Weka and clustered using the *k*-Means algorithm. At first, we performed the clustering multiple times with a different number of clusters each time to determine the optimal number.

It was decided there should be 4 clusters. It was observed that every additional cluster just seemed to provide nearly the same data as one of the existing clusters. Clusters will be labelled *A*, *B*, *C* and *D*. Cluster A would point towards the best cluster, and Cluster D would point towards the worst cluster.

5.2.1 PITUP Križevci

Table 4 shows the number of students for each cluster of PITUP Križevci dataset as well as the percentage they take from the total number of students for this centre. Table 5 shows the centroids for each cluster dataset. This dataset has the least number of students in all datasets.

Table 4. Number and percentage of students per cluster from PITUP Križevci

Cluster	Count	%
A	5	15.15%
B	5	15.15%
C	8	24.24%
D	15	45.45%
Total	33	100.00%

Table 5. Centroids for PITUP Križevci

Attribute	Clusters			
	A	B	C	D
File	12.00	6.80	4.75	2.13
Forum	13.80	9.80	11.75	5.67
Student report	4.80	4.00	3.88	0.40
Lesson	1979.2	1625.8	840.00	211.73
File upload	0.00	0.0	0.00	0.00
Link	0.40	0.20	0.75	0.47
Overview report	0.20	0.00	0.38	0.00
Page	2.60	2.00	3.00	1.67
System	359.60	271.00	145.13	43.93
Test	448.80	237.20	54.00	0.53
Home-work	0.00	0.00	0.00	0.00
SA 1	0.81	0.70	0.52	0.00
SA 2	0.89	0.81	0.08	0.00
SA 3	0.83	0.73	0.00	0.00
SA 4	0.82	0.53	0.00	0.00
SA 5	0.82	0.52	0.00	0.00
SA 7	0.87	0.00	0.00	0.00
SA 8	0.80	0.40	0.21	0.00
SA 9	0.43	0.0	0.09	0.00
SA Average	0.70	0.41	0.10	0.00

As it is shown from the data in the presented tables, we can see that only a small number of students (clusters A and B) actively took self-assessment tests while the rest (clusters C and D) tried a few times or did not try at all.

Students from cluster A had a clear lead in their activity on Moodle LMS and their success on self-assessments tests. It is worth noting that *SA 9* seems to be ruining their *SA Average*.

Students from cluster B were almost as active as cluster A, but their average success on self-assessment tests is mediocre at best. They tried their best on the first three self-assessment tests, but slowly stopped trying for the other five.

Students from cluster C weren't as active as students from clusters A and B, but are still more active than the students from cluster D. However, their success on self-assessment tests were quite low. They tried with the first test but seemed to give up on every other test except for self-assessment test 8 (*SA* 8).

Students from cluster D were the least active students and they haven't even tried taking the self-assessment tests. Students from this cluster were also in the majority of this PITUP centre. In numbers, 15 out of 33 students from PITUP Križevci dataset had no interest in taking self-assessment tests.

5.2.2 PITUP Sisak

Table 6 shows the number of students for each cluster for PITUP Sisak, as well as the percentage they take from the total number of students from this centre. Table 7 shows the cluster centroids. This dataset has second lowest number of students in all datasets

Table 6. Number and percentage of students per cluster from PITUP Sisak

Cluster	Count	%
A	12	16.22%
B	13	17.57%
C	13	17.57%
D	36	48.65%
Total	74	100.00%

Table 7. Centroids for PITUP Sisak

Attribute	Clusters			
	A	B	C	D
File	17.42	10.31	8.69	2.61
Forum	49.33	9.00	10.23	7.75
Student report	10.75	3.31	3.85	0.97
Lesson	3226.6	2448.6	1242.4	190.28
File upload	0.00	0.00	0.00	0.00
Link	0.67	0.23	0.85	0.17
Overview report	0.42	0.00	0.00	0.00
Page	3.92	1.77	2.46	0.56
System	638.67	295.69	179.77	49.36
Test	764.58	264.85	53.85	0.08
Home-work	0.00	0.00	0.00	0.00
SA 1	0.81	0.69	0.47	0.00
SA 2	0.85	0.83	0.32	0.00
SA 3	0.84	0.83	0.05	0.00
SA 4	0.81	0.54	0.00	0.00
SA 5	0.87	0.47	0.00	0.00
SA 7	0.91	0.00	0.00	0.00
SA 8	0.73	0.53	0.04	0.00
SA 9	0.70	0.16	0.00	0.00
SA Average	0.72	0.45	0.10	0.00

The results indicated that students from cluster A were most active and most successful on the self-assessment tests. An interesting thing to note was the value of their attribute *Test*. It was almost 3 times bigger than the value in cluster B, which was the second largest value of all clusters. A possible interpretation could be that students have been continuously retaking the self-assessment tests until they had a high enough score, either as a proof to themselves that they learned the specific topic or simply as a type of self-accomplishment.

Students from cluster B were somewhat less active than the students from cluster A, but they still had some activity. As for their success in self-assessment tests, they were mediocre at best. Similarly to PITUP Križevci, they started strong but gradually receded.

Students from cluster C had a decent activity. They were more active than their PITUP Križevci counterparts. However, their success on self-assessment tests was weak.

Students from cluster D were the least active. From their success on self-assessment tests one could assume they haven't given those tests any thought. Almost half of students from PITUP Sisak dataset belonged to this cluster.

5.2.3 PITUP Zabok

For PITUP Zabok dataset, Table 8 shows the number of students for each cluster as well as the percentage they take from the total number of students from this centre, and Table 9 shows the centroids for the computed clusters. This dataset has the second greatest number of students in all datasets, only one less than PITUP Varaždin dataset.

Table 8. Number and percentage of students per cluster from PITUP Zabok

Cluster	Count	%
A	11	8.87%

B	15	12.10%
C	14	11.29%
D	84	67.74%
Total	124	100.00%

Table 9. Centroids for PITUP Zabok

Attribute	Clusters			
	A	B	C	D
File	11.91	9.67	12.29	4.50
Forum	30.45	14.20	19.43	8.75
Student report	4.73	3.13	4.8571	1.21
Lesson	2591.3	2210.1	1745.1	243.85
File upload	0.00	0.00	0.00	0.00
Link	0.45	0.67	1.21	0.29
Overview report	0.36	0.27	0.21	0.06
Page	0.91	1.40	2.21	1.00
System	361.45	277.00	268.71	62.54
Test	422.36	257.67	91.57	2.58
Home-work	0.00	0.00	0.00	0.00
SA 1	0.76	0.68	0.51	0.03
SA 2	0.82	0.87	0.53	0.00
SA 3	0.84	0.84	0.24	0.00
SA 4	0.82	0.69	0.03	0.00
SA 5	0.83	0.68	0.00	0.00
SA 7	0.71	0.06	0.00	0.00
SA 8	0.80	0.42	0.28	0.00
SA 9	0.52	0.10	0.00	0.00
SA Average	0.68	0.48	0.18	0.00

Again, students from cluster A had the most activity, but not by a wide margin when comparing them to students from cluster B. They were also most successful on self-assessment tests. Like in PITUP Križevci dataset, their *SA Average* was slightly ruined by *SA 9*.

Students from cluster B were slightly less active than students from cluster A, but their average success on self-assessment tests were mediocre just like in previous datasets.

Cluster C students had a quite decent activity, slightly less than students from cluster B. Their success on self-assessment tests is quite low but it can be seen that students tried in the first two test and gave up on the other ones.

Lastly, students from cluster D had the least activity and made almost no attempts to solve the self-assessment tests. Also, over half of students from PITUP Zabok dataset belonged to this cluster, which was much greater than in other datasets.

5.2.4 PITUP Varaždin

Table 10 shows number of students per cluster and the percentage they take from the total number of students from this centre. Table 11 shows the cluster centroids for this dataset. This dataset has the greatest number of students in all datasets.

Table 10. Number and percentage of students per cluster from PITUP Sisak

Cluster	Count	%
A	35	28.00%
B	26	20.80%
C	30	24.00%
D	34	27.20%
Total	125	100.00%

Table 11. Centroids for PITUP Varaždin

Attribute	Clusters			
	A	B	C	D
File	25.60	18.38	21.63	6.65
Forum	63.37	49.69	55.03	27.12
Student report	24.43	9.15	13.30	3.88
Lesson	3083.3	2716.9	1591.7	424.82
File upload	4.69	6.62	7.13	2.94
Link	1.17	0.81	1.33	0.79
Overview report	0.49	0.19	0.17	0.03
Page	2.63	1.62	2.63	0.88
System	572.63	406.23	373.23	111.79
Test	470.46	325.12	122.77	18.65
Home-work	35.37	34.92	42.40	15.03
SA 1	0.80	0.72	0.56	0.10
SA 2	0.91	0.84	0.41	0.00
SA 3	0.89	0.83	0.16	0.00
SA 4	0.82	0.58	0.00	0.00
SA 5	0.87	0.39	0.00	0.00
SA 7	0.90	0.00	0.00	0.00
SA 8	0.80	0.47	0.13	0.00
SA 9	0.68	0.21	0.03	0.00
SA Average	0.74	0.45	0.14	0.01

Cluster A contains the most active students which were also most successful on self-assessment tests.

Similar to previous datasets, the value in attribute *SA 9* is slightly ruining the value of attribute *SA Attribute*.

Students from cluster B were only slightly less active comparing to students from cluster A. But, as it was the case with B clusters in previous datasets this cluster also had a mediocre average success on self-assessment tests. Students from cluster B seemed to be the students that start strongly in first few tests, but then stop giving much effort into other ones.

While students from cluster C had the second least activity, they were still decently active. However, their successes on the self-assessment tests were weak.

As for the students from cluster D, it's interesting to note that they were the most active when comparing them to cluster D students from other datasets. But they still seemed to not give much effort to taking self-assessment tests as their successes on them were mostly zero.

What's even more interesting is that there were the same number of students belonging to clusters A and D. In previous datasets, there were just slightly-under-half or slightly-over-half of total number of students belonging to cluster D. Even PITUP Zabok datasets, which had the same number of students, had 50 more students in cluster D than this dataset.

6 Conclusion

From the data presented in this paper and its interpretation we can conclude that the more active a student is on the LMS the more likely he/she is going to take the self-assessment test with success. Also, with the cluster analysis we concluded that in PITUP centres Križevci, Sisak and Zabok most of the students are profiled as non-active students. This could mean that students' activities and performances on the self-assessment tests are also influenced by location and/or different studying terms in PITUP centres outside Varaždin.

However, we need to emphasize that students in Varaždin are given also live classes of 30 hours (2 hours per week), and are mostly full-time students. However, students from other three centres are all part-time students, most of them travel to get to classes, and for them traditional classes are held only twice (5 hours each).

Moreover, students in Varaždin are given scores for their activity in LMS and other activities in the course throughout the semester, while students in other centres are not rewarded for any of the extra activities (including self-assessment). Instead, the self-assessment tests are just a feedback mechanism that help them to prepare for the final exam.

There are several possible directions for future research. We could take students' performances on the self-assessment tests and their course grades to determine if the former influences the latter. Cluster analysis could also be applied for the next academic year, or next several academic years, and compare them to see if the next generations of students will be more, less, or equally active on LMS and successful on self-assessment tests. This would be beneficial in case of PITUP centres where less active students are a majority (Križevci, Sisak and Zabok).

7 References

Asif R., Merceron A., Ali S.A. & Haider N.G., Analyzing undergraduate students' performance using educational data mining, *Computers & Education* (2017), doi:10.1016/j.compedu.2017.05.007

Baker R.S.J.d. (2010). Data Mining. In McGaw, B., Peterson, P., Baker, E. (Eds.) *International Encyclopaedia of Education (3rd edition)*. Oxford, UK: Elsevier.

Bouchet, Harley, J., Trevors, G., & Azevedo, R. (2013). Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning. JEDM | Journal of Educational Data Mining, 5(1), 104-146. Retrieved from https://jedm.educationaldatamining.org/index.php/JEDM/article/view/32

Bovo, A., Sanchez, S., Héguy, O., & Duthen, Y. (2013, September). Clustering moodle data as a tool for profiling students. In e-Learning and e-Technologies in Education (ICEEE), 2013 Second International Conference on (pp. 121-126). IEEE.

Cantabella M., Martínez-España R., Ayuso B., Yáñez J. A., Muñoz A. (2018). Analysis of student behavior in learning management systems through a Big Data framework, *Future Generation Computer Systems*, 90, Pages 262-272, doi: https://doi.org/10.1016/j.future.2018.08.003.

Ervina Alfan, Md Nor Othman, (2005) Undergraduate students' performance: the case of University of Malaya, Quality Assurance in Education, Vol. 13 Issue: 4, pp.329-343, https://doi.org/10.1108/09684880510626593

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. The Internet and Higher Education, 28, 68-84.

Kadoić, N., & Oreški, D. (2018, January). Analysis of Student Behaviour and Success Based on Logs in Moodle. In 41st International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO 2018.

Pislaru, C., & Mishra, R. (2009, April). Using VLEs to support student centred learning in Control Engineering Education. In Proc. 5th Int. Conf on Multimedia and Information and Communication

Technologies (m-ICTE 2009), University of Lisbon, Portugal (pp. 22-24).

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33, 135-146.

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. Computers & Education, 68, 458-472.

Rosalina Rebucas Estacio, Rodolfo Callanta Raga Jr, (2017) "Analyzing students online learning behavior in blended courses using Moodle", *Asian Association of Open Universities Journal, Vol. 12* Issue: 1, pp.52-68, https://doi.org/10.1108/AAOUJ-01-2017-0016

Saarela, M., & Kärkkäinen, T. (2015). Analysing Student Performance using Sparse Data of Core Bachelor Courses. JEDM | Journal of Educational Data Mining, 7(1), 3-32. Retrieved from https://jedm.educationaldatamining.org/index.php/JEDM/article/view/JEDM056

Talavera, L., & Gaudioso, E. (2004). Mining Student Data To Characterize Similar Behaviour Groups In Unstructured Collaboration Spaces. In *Workshop on artificial intelligence in CSCL. 16th European conference on artificial intelligence* (pp. 17–23).

The University of Waikato (2017). Weka (3.8.2.) [Data mining software]. Retrieved from https://www.cs.waikato.ac.nz/ml/weka/

Wang Jie, Lv Hai-yan, Cao Biao and Zhao Yuan, Application of educational data mining on analysis of students' online learning behavior, *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, Chengdu, 2017, pp. 1011-1015. doi: 10.1109/ICIVC.2017.7984707. Retrieved from https://ieeexplore.ieee.org/document/7984707/