Student Modeling with Clustering: Comparative Analysis of Case Studies in Two Higher Educational Institutions from Different Countries

Dijana Oreški [1], Emilija Kisić[2], Jovana Jović[2] and Miroslava Raspopović Milić[2]

[1] *University of Zagreb, Faculty of Organization and Informatics, Pavlinska 2, 42000 Varaždin, Croatia*
[2] *Belgrade Metropolitan University, Tadeuša Košćuška 63, 11000 Belgrade, Serbia*

Abstract

The focus of this paper is to combine learner data from two different higher education institutions, and specifically, from one course from each institution. Data were collected during one semester in both courses and stored in historical datasets with the aim to identify the most important features from each dataset that contribute to an adequate grouping of students based on their pre-exam activities and gained knowledge during the course's duration. The goal of this paper is to analyze learner data and compare the results with the lesson designs of both courses. The aim is to address future needs for the lesson designs that will be used in a system with lesson content recommendations based on student modeling used for tracking of student-specific progress.

Keywords

Student modeling, clustering, lesson design, student progress tracking

1. Introduction

An important part of intelligent tutoring systems is a student model [1]. The student model should represent the student's current state of knowledge and abstraction of a student's characteristics, behaviors, learning preferences, and performance [2], [3]. These models are essential in learning environments because they allow personalization of the learning process for each student. Besides the personalization of the learning content, student models also offer the opportunity for tailored interventions and focused support when necessary [4], [5].

Through the analysis of huge amounts of data gathered from students' interactions with learning platforms, resources, and tests, artificial intelligence (AI) can be used for student modeling [6]. AI models use complex algorithms to interpret student's habits, preferences, learning styles, etc. By analyzing learner data, AI creates student models that student's unique learning preferences, skills, and knowledge gaps [7], [8]. Student modeling can be used to capture student characteristics such as knowledge level, learning style, strengths, shortcomings, and progress trajectory, and in some cases possibly predict them [9], [10]. In order to offer personalized learning experiences it is important to identify trends in learner data. Using clustering techniques is one approach that can allow to identify these trends [11]. Learning patterns and behaviors shared by students are the basis for clustering algorithms, which divide students into discrete categories or clusters. Teachers can use generated clusters to understand and meet the requirements of a wide range of students [12]. Moreover, clustering is an effective technique that identifies groups of students with similar characteristics. This allows for tailored content recommendations and focused interventions for each student cluster [13], [14].

Proceedings for the 14th International Conference on e-Learning 2023, September 28-29, 2023, Belgrade, Serbia
EMAIL: dijana.oreski@foi.hr (A. 1); emilija.kisic@metropolitan.ac.rs (A. 2); jovana.jovic@metropolitan.ac.rs (A. 3); miroslava.raspopovic@metropolitan.ac.rs (A. 4)
ORCID:0000-0002-3820-0126 (A. 1); 0000-0003-3059-2353 (A. 2) ; 0000-0002-4204-0233 (A. 3); 0000-0003-2158-8707 (A. 4)

In order to effectively develop student models based on learner data, lesson design with planned learner activities (quizzes, tests, etc.) should be carefully planned and well-structured [15]. A setting that will allow for easier tracking of student progress includes planning classes with specific goals, designing learning activities that will engage students, and hence, allow for collection of quality learner data, and finally, plan adequate assessments [16]. Points in the lesson where learner data that allows student progress tracking is collected are called lesson breakpoints. Breakpoints can be in the form of assessment, progress report submission, etc. Even though many different forms of assessments can be implemented in a lesson, they can all serve a purpose of collecting measurable progress that can show how engaged, knowledgeable, and successful the students are in the learning at the specific course [17]. In addition, gathering good quality quantitative and qualitative information will allow for building more precise student models.

In this work, learner data from two different universities were collected and analyzed. The aim was to identify patterns in data utilizing clustering algorithms and use the identified clusters to identify students with different progress levels in their learning. Additionally, the idea was to determine key features in both datasets and use the conclusions for future lesson designs suitable for more precise tracking of student progress [18]. By doing this, institutions want to improve the educational experiences of their students, and at the same time make a significant contribution to the wider conversation about the potential of artificial intelligence in adaptive e-learning and how to use that potential through adequate lesson design and process data more easily. This research has the following goals: (i) to compare and analyze the learner data and used lesson designs from two different courses to identify characteristics of similar clusters of students based on their activity levels, and (ii) to use these findings to address needs for the future lesson design needed for learning content recommender systems based on student modeling. This paper is organized as follows: Section 2 describes the methodology used for collecting and analyzing data in both involved institutions. Section 2.1 within section 2 presents lesson designs used in analyzed courses. Section 2.2 within section 2 describes characteristics and features of both learner datasets. Section 3 presents the results and points out key findings about the conducted clustering and feature importance analysis. Finally, Section 4 concludes the paper.

2. Methodology

In this paper, two courses from two higher education institutions were analyzed: Belgrade Metropolitan University, Faculty of Information Technology (FIT, BMU), Serbia, and University of Zagreb, Faculty of Organization and Informatics (FOI, UNIZG), Croatia. For comparative analysis, two courses with lesson design shown in Figure 1 were chosen from each institution. Course "Discrete Structures" was selected from BMU and the course "Informatics" was selected from FOI. Learner data that were collected during a semester from both courses were gathered from institutions' Learning Management Systems (LMSs). The main goal of this work was to analyze student activities in the course and group students based on similar features into groups. The aim was to identify which features give high importance for adequate student grouping. The identification of key features for modeling student progress is useful for addressing necessary improvements for the future lesson designs that will be most suitable for extracting useful information about student engagement and progress during the course duration. The idea is for the future lesson design to include student modeling for tracking student progress and based on the finding on the level of the progress to recommend student learning content. In order to do so, in two different institutions, course context and lesson designs from both institutions were compared, analyzed and intersected with learner data information, in order to make recommendations for future lesson designs. On its own, student model achieves very little, but with the use of artificial intelligence techniques, such as clustering and recommendation techniques, it can be useful for real time tracking of student-specific progress.

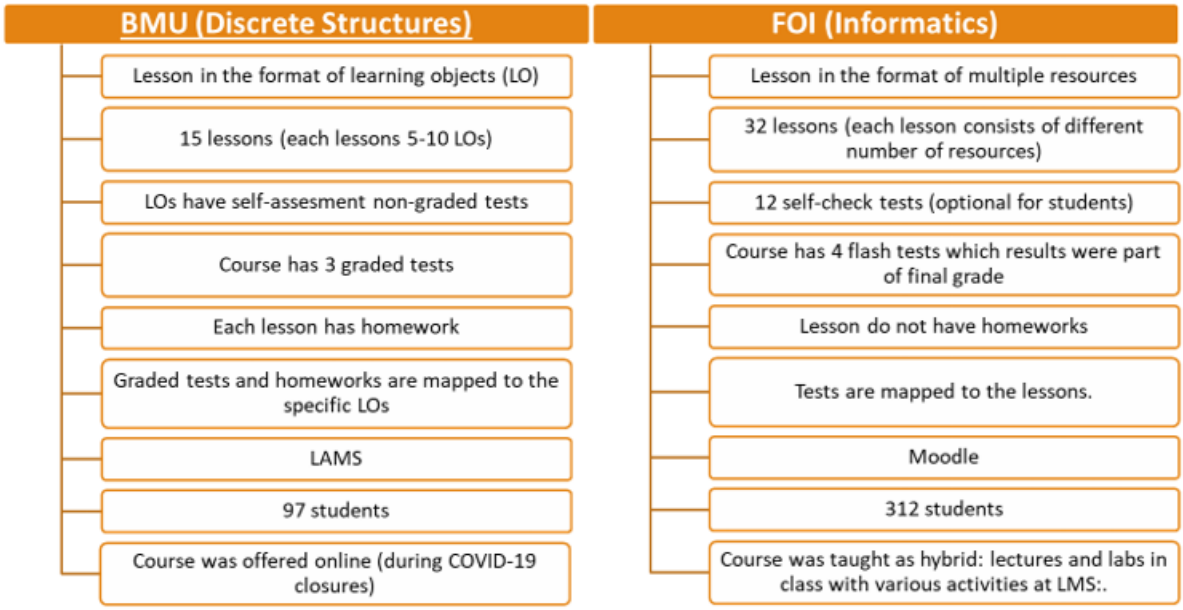

Figure 1: Comparison of key components of lesson designs for BMU and FOI

In order to form datasets for both courses different LMSs were used at each institution. On both courses, different features for forming of datasets were chosen. After datasets were formed, grouping of students was performed using K-means clustering [19]. After clusterization, feature importance was performed for gaining the most important features that contribute to the student grouping. After gaining the results on both institutions, the results were compared, and conclusions were made for better future lesson design in order to gain reliable student modeling and accurate recommendations for real time tracking of student-specific progress.

2.1 Selected lesson designs

BMU's course, Discrete structures, was attended by 97 students. This course is offered to first and second-year students studying in undergraduate academic programs - Information Technology, Software Engineering and Video Games Development. The particular instance of the course, that is a part of this analysis, was taken from the course taught during the COVID-19 pandemic and school closures. Teaching and learning materials were posted on the Learning Activity Management System (LAMS), while students had weekly lectures with a professor over the Zoom platform. Teaching and learning materials posted on LAMS were structured as a combination of a series of learning objects (LOs) and lesson activities. Each lesson consisted of a set of LOs (5-10 LOs). Besides the instructional materials, lessons contained activities such as not-graded self-assessments after each LO, forum, shared resources, and mind maps. An example of BMU's lesson design with LOs and learning activities is presented in Figure 2. Besides the non-graded lesson activities, students were also given graded assignments such as homework assignments each week, and approximately every 4-5 weeks, students were given a test. Both tests and homework assignments were mapped to a specific LO. In other words, each homework and test evaluated particular knowledge from a LO or set of LOs. Furthermore, tests were designed with five sections, each section was also mapped to a specific LO or set of LOs. This was done with an intention, to be able to easier point students towards the recommended material, should they not pass homework or tests with minimum grade. An example of test organization with sections is shown in Figure 3. The lesson design, which included mapping teaching materials and formative assessments for specific sections, facilitated easier tracking of student progress and identification of areas where students may be struggling within the course.

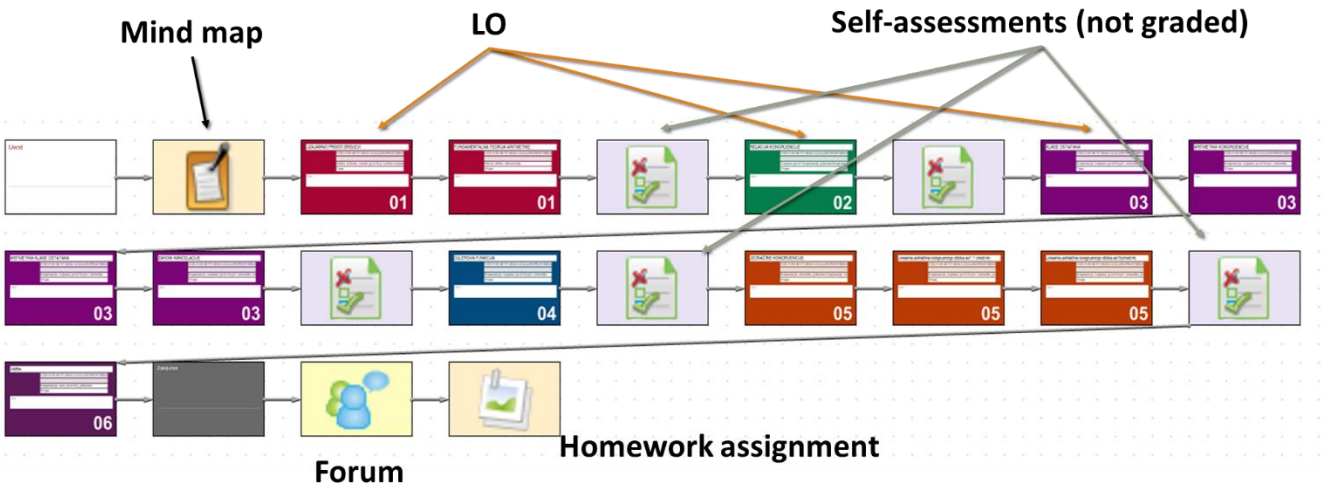

Figure 2: Example of BMU's lesson design with learning objects and lesson learning activities

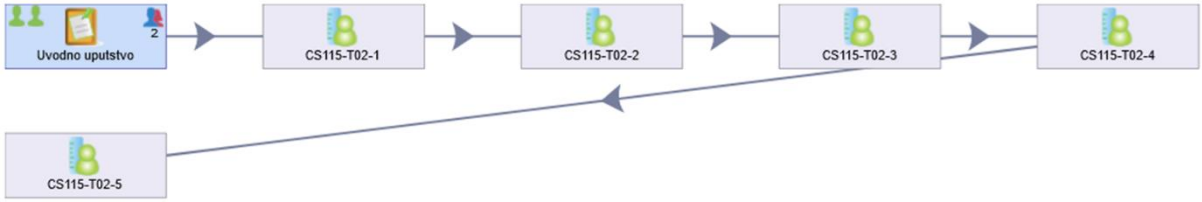

Figure 3: Example of BMU's organization of graded tests into different sections

FOI's course Informatics was taught as mandatory in the first semester of the undergraduate study program Information and Business Systems at the University of Zagreb's Faculty of Organization and Informatics, and 312 students enrolled in the course. The course was held in a hybrid format, consisting of lectures and laboratory exercises at the faculty along with course materials and assignments available through LMS Moodle. Lecture topics were presented to students at the LMS through lesson activities. Overall, 32 lessons were created on Moodle. Lessons were in multi-resource format, consisting of text, pictures, videos, and links. Students' knowledge evaluation, among others, was performed by using self-assessment tests and flash tests.

There were 12 self-assessment tests throughout the semester. Self-assessment tests were optional, meaning their results were not part of the final grade. Students could assess it in order to test their knowledge and as preparation for the midterm exam. Flash tests, on the other hand, were mandatory, and their results were part of the student's final grade. Questions at self-assessment tests and flash tests were mapped to the lesson contents. Example of the lesson organization is shown in Figure 4.

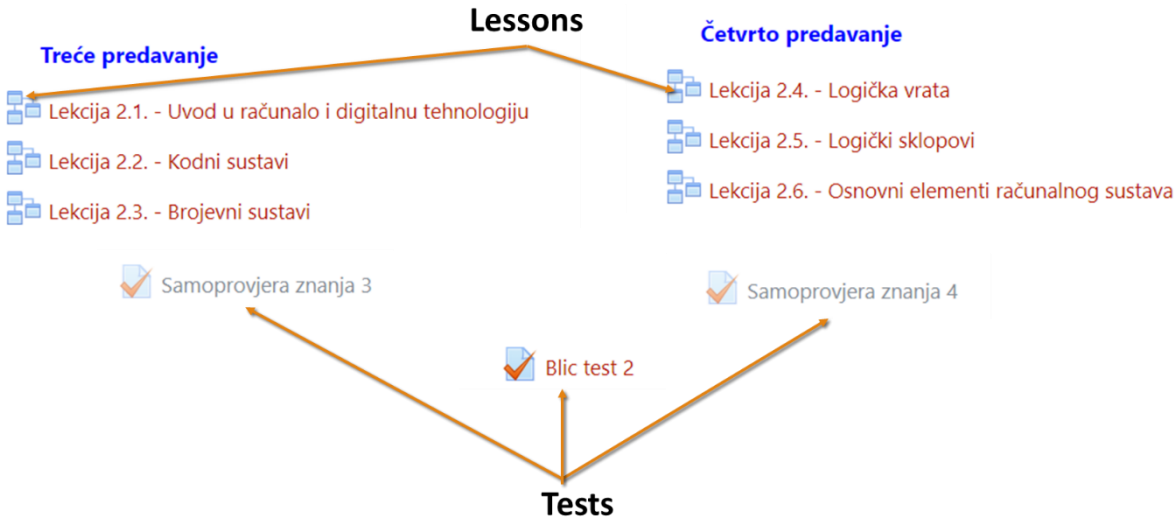

Figure 4: Example of FOI's lesson design

Typically, one course topic was covered with three distinct lesson sections. The lecture entitled *Treće predavanje* consisted of lessons: *Lekcija 2.1*, *Lekcija 2.2*, and *Lekcija 2.3*. Self-assessment test entitled *Samporocjena znanja 3*, corresponds to a lecture *Treće predavanje*. Furthermore, the flash test identified as *Blic test 2* assessed gained knowledge from two lectures: *Treće predavanje* and *Četvrto predavanje*, and total of six lessons: *Lekcija 2.1, Lekcija 2.2, Lekcija 2.3, Lekcija 2.4, Lekcija 2.5*, and *Lekcija 2.6.*

A comparison of key components of lesson designs for BMU and FOI is shown in Figure 1. As we can see from Figure 1, there are several differences between selected courses in lesson design. Beside different LMSs and course organizations, one of the key differences is lesson granulation on selected courses and mapping of student activities on specific parts of the lesson or at the lesson as a whole. At BMU lessons are represented through LOs and student activities are mapped to specific LOs, while at FOI there is no mapping to a specific part of the lesson; there is mapping to the lesson as a whole. That will lead to differences in datasets in the sense that in BMU dataset student activities are mapped on each LO within the lesson, while at FOI student activities refer to the entire lesson, without mapping to specific part of the lesson. This means that learner data for BMU in the dataset is represented through LOs, while for FOI's learner data through the lessons.

2.2 Data description

The dataset obtained from BMU is based on students' activities during the semester. This dataset contains four features:

1. 'Test points': This variable represents the number of points received on a graded test.
2. 'Homework points': This variable represents the number of points received on a homework assignment.
3. 'Assessment points': This variable represents the number of points received on non-graded self-assessment tests.
4. 'Time spent on each LO': This variable was formed as a column with a true or false value, where true means that student spent some time on specific LO, while false means that student did not spend any time on specific LO.

FOI dataset consists of data about students' LMS Moodle activities, focusing on the lessons. FOI's dataset includes following variables:

1. 'Lesson is initiated: This variable represents the number of logs for the lesson, saying how many times a student started the lesson.
2. 'Lesson is completed': This variable represents the number of logs for each time a student has viewed the entire lesson.
3. 'Lesson is continued': This variable represents the number of logs for each time when a student has returned to the already initiated lesson.
4. 'Lesson is restarted': This variable represents the number of logs when the student started the lesson again.
5. 'Lesson has been reviewed': This variable represents the number of logs in which students revisited and reopened the lesson again.
6. 'Self-assessment test points': This variable provides information about the number of points that students achieved in self-assessment tests.
7. 'Flash test points': This variable provides information about the number of points that students achieved in flash tests.

The first five variables are related to students' lessons viewing frequency, whereas the last two variables, self-assessment test points and flash test points, provide insight into their performance. Features for both datasets are shown in Table 1.

Table 1
Data description

FIT BMU data	FOI UNIZG data
Test points.	Lesson is initiated.
Homework points.	Lesson is completed.
Assessment points.	Lesson is continued.
Time spent on each LO.	Lesson is restarted.
	Lesson has been reviewed.
	Self-assessment test points.
	Flash test points.

It can be seen from Table 1 that datasets have different features, but similar in the sense that selected course features describe student's 'time engagement' on specific lesson or part of the lesson, and obtained points on graded or non-graded assessments and tests. Differences among features and different lessons granularity will lead to diverse conclusions and will give better insight for the future lesson design.

3. Results

In order to group students based on their activities, course progress during the semester, and obtained knowledge on pre-exam activities, K-means clustering was performed [19]. Different numbers of clusters were varied for both datasets. For the presentation of the results in this paper, three clusters were selected, for the purpose of easier comparison between features of two different datasets. Grouping the students in two clusters was not considered for this paper, as the simplification of two clusters (engaged and not engaged) would simplify the analysis too much. Clustering identified clearly three groups of students: very active learners, moderately active learners, and passive learners. Students could also be grouped into more than three clusters which will lead to more precise grouping and engagement distinguishing, which will be considered for future work.

Significance of being able to distinguish between different levels of course progression is important for the future work, where the goal is to recommend learning content to each student based on their activity level and obtained knowledge. With the clustering technique, datasets can be labeled and labels can be used for specific parts of the lesson recommendation or for lesson recommendation as a whole. Considering identified activity level groups, if a student is in the group of very active learners that means that based on the student's pre-exam activities the student gained enough knowledge. In the future student modeling context, if the student satisfied the progress level and obtained knowledge, that student will not be recommended to revisit a certain part of the lesson. If a student is in the group of

moderately active learners that means that student most likely did not acquire enough knowledge and is not progressing fully in the course. This student will be recommended to revisit certain parts of the lesson. If a student is in the group of passive learners, that means that student did not gain enough knowledge and is lacking in the course activities. This would mean that the future recommender system will most likely point this student to revisit certain parts or the entire lesson. In this way, each student could have a recommendation for whether he should relearn a specific part of the lesson, and whether he has to learn it in detail. This case study was performed on historical datasets, however in real time, with this type of recommendations in the future, it will be possible to have specific-student progress tracking during the semester and to intervene with feedback to the student at certain breakpoints for his better progress.

As an example of conducted clustering, Figure 5 shows grouping into three clusters for BMU data. For better visualization, dimension reduction was performed and that is presented with three dimensions for better visualization. Dimensionality reduction is performed with PCA analysis [20]. In Figure 5 it can be seen that clusters are well separated, meaning that selected features make significant differences between student groups.

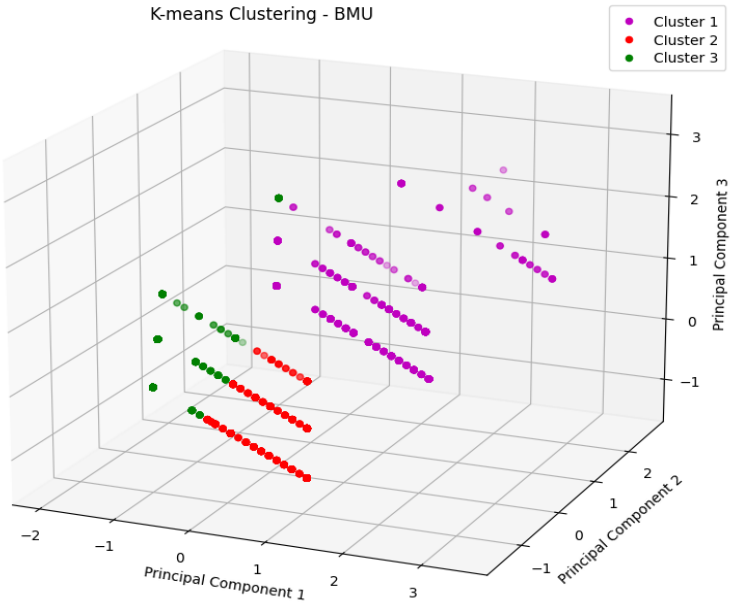

Figure 5: K-means clustering for BMU students after dimensionality reduction.

Percentages of students in each of three groups, after clustering in BMU and FOI, are shown in Figure 6. Within BMU's dataset, the largest identified group is a group of very active learners, and the smallest group are passive learners. On the other hand, more than half of students from FOI's dataset were classified under the passive learners. About a third of the FOI students are shown to be very active learners. As expected, the distribution of student numbers for the clusters are different for both sets, as the context of teaching is different. Also, it is interesting to address the difference in the lesson designs at this stage, as the content granularity between two courses is different. At BMU, lessons are made of LOs and groups are based on activities which are mapped to specific LOs within the lessons, while at FOI content is at the lesson level, without further granularity, and therefore, groups are based on activities which refer to the lesson. Thus, at BMU for a specific student, gained knowledge and engagement refer to each LO within the lesson, which explains the dominant group of active learners. On the other hand, at FOI, gained knowledge and engagement refer to the lesson as a whole, so the dominant group are passive learners. After clustering and comparative analysis, we can conclude that in future lesson design it is desirable to have larger lesson granularity for gaining more precise learning content recommendation.

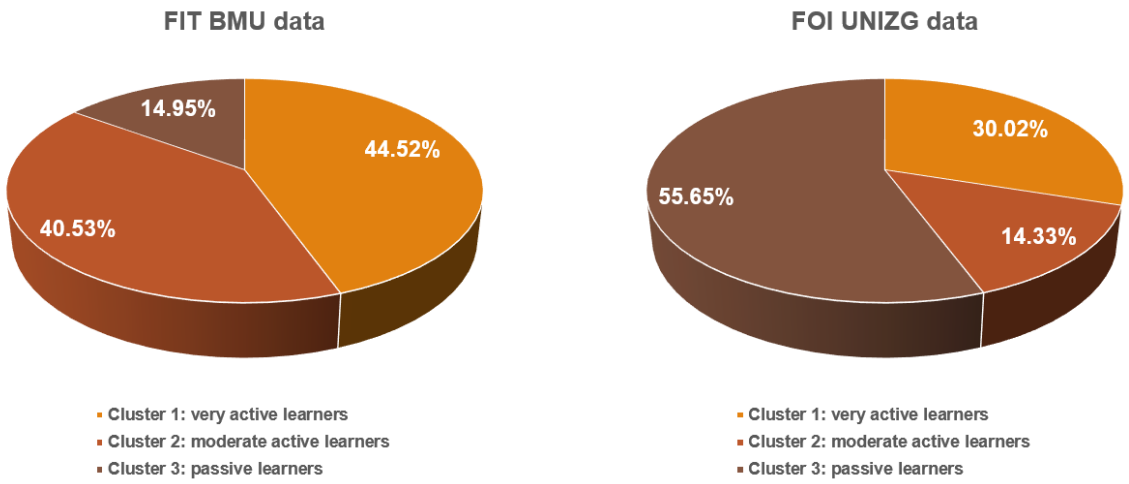

Figure 6: Distribution of students into three groups after clustering at BMU and FOI.

For further analysis, feature importance analysis was conducted in order to depict variables with a higher contribution to the cluster differentiation. Feature importance analysis was undertaken for both datasets. For the BMU dataset the most important discriminator is the feature 'Test points.' This was expected, because among other features this one was the most informative concerning a student's level of knowledge and engagement during the semester. Regarding the temporal nature of BMU data, the course was held during Covid pandemic, so all tests and homework assignments were taken from home, only the final exam was taken in the classroom. This is something to keep in mind, because of the objectivity of the received points on the take-at-home tests.

Within the FOI dataset, the following variables emerged as the most pivotal discriminators: 'Self-assessment test points' and 'Lesson is initiated.' The data analysis conducted on the FOI dataset revealed a high level of temporal pattern. There is a noticeable trend characterized with increased student activity at the beginning of semester and after that these activities decreased during the middle of semester. Another increase in student activity was noticed after the end of the first midterm exam. This could be attributed to a change in students' motivation during the semester which was not big at the start of semester and it increased after their dissatisfaction with grades on the midterm exam. FOI dataset temporal pattern should be considered for the future lesson design. In the period of decreased activities and motivation some additional activities should be undertaken to overcome this problem.

Considering that in both institutions key discriminators were identified, they should be included in future lesson designs. Considering that test points are showing as a common feature, it should be a mandatory feature for future student modeling. It may not be obvious how to include "lesson is initiated" in the lesson design, however, this can be thought of in terms of planning student active engagement and it is recommended from these results to include this feature or some similar features in the future data collection.

4. Conclusions and future work

This paper analyzed learner data from two different higher educational institutions. One course was taken from each institution. The goal of this analysis was to identify key features that contribute to adequate grouping of students in several groups based on their pre-exam activities and gained knowledge during the course's duration. K-means clustering was used to separate students into three clusters: very active learners, moderately active learners, and passive learners. Using feature importance of data, graded assessments and tests were identified as the key discriminators. Conclusions after comparative analysis gave insight into more appropriate lesson design planning for the future courses where student modeling based recommendations are planned. In other words, formative testing should be planned in the course as soon as possible, in order to be able to track student progress as early as possible. The temporal nature of the BMU's data should be taken into consideration. BMU's course was held during the pandemic, and students completed their assessments and tests from home. On the other hand, the temporal pattern in the FOI dataset revealed fluctuations in student activity and motivation throughout

the semester. This suggests that needed interventions and course activities should be aware of these temporal instances and should keep engaging and motivating students more as the semester progresses.

In the lesson design of each institution different granularity levels of teaching materials were used. BMU used learning objects as their sections of the lessons, and mapped their assessments and assignments to particular LOs. FOI used lesson formats without finer mapping to the lesson sections. After the clustering, it was evident that different granularity produced different distribution of students among clusters and we can conclude that finer granularity of teaching content in future will lead to more precise recommendations of learning content. Having in mind that graded assessments and tests showed the highest significance in feature importance, this pointed towards the need to carefully plan their placement in the lesson design. It is recommended that graded assignments are placed as early as possible, where they will be used to assess certain knowledge or learning outcomes. Assessments should also be designed with a level of granularity that allows for the pinpointing parts of the lesson that the student has not fully adopted. Future lesson design should additionally consider incorporating elements that will provide a greater number of features for machine learning algorithms. This includes considering elements such as the frequency of student access to each section, completion status, test points, and homework points. This comprehensive approach will enable a more diverse feature collection and successful data integration between institutions with the final goal of real time specific-student progress tracking and learning content recommendations.

Acknowledgements

The paper is co-supported by the Ministry of Science, Technological Development and Innovations of the Republic of Serbia ref. no. 451-03-47/2023-01/200029.

References

[1] H. L. Burns and C. G. Capps, “Intelligent tutoring systems: an introduction,” *Foundations of intelligent tutoring systems*, vol. 1, 2013.

[2] P. Sedlmeier, “Intelligent Tutoring Systems,” in *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier, 2001, pp. 7674–7678.

[3] E. Mousavinasab, N. Zarifsanaiey, S. R. Niakan Kalhori, M. Rakhshan, L. Keikha, and M. Ghazi Saeedi, “Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods,” *Interact. Learn. Environ.*, vol. 29, no. 1, pp. 142–163, Jan. 2021.

[4] A. Keleş, R. Ocak, A. Keleş, and A. Gülcü, “ZOSMAT: Web-based intelligent tutoring system for teaching–learning process,” *Expert Syst. Appl.*, vol. 36, no. 2, pp. 1229–1239, Mar. 2009.

[5] A. Shemshack and J. M. Spector, “A systematic literature review of personalized learning terms,” *Smart Learning Environment*, vol. 7, no. 1, Dec. 2020.

[6] K. Seo, J. Tang, I. Roll, S. Fels, and D. Yoon, “The impact of artificial intelligence on learner-instructor interaction in online learning,” *International Journal of Educational Technology in Higher Education*, vol. 18, no. 1, p. 54, Oct. 2021.

[7] M. Ilić, V. Mikić, L. Kopanja, and B. Vesin, “Intelligent techniques in e-learning: a literature review,” *Artificial Intelligence Review*, Jun. 2023.

[8] F. Kamalov, D. Santandreu Calonge, and I. Gurrib, “New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution, ” Sustain,” *Sustain. Sci. Pract. Policy*, vol. 15, no. 16, 2023.

[9] J. Jovic, D. Domazet, M. Milic, and K. Chandra, “Student model in inteligent tutoring systems - a systematic review,” in *Proceedings of the 11th International Conference on eLearning (eLearning-2020)*, Belgrade, Serbia, 2020, pp. 24–25.

[10] J. Jović, E. Kisić, M. Raspopović MIlić, D. Domazet, and K. Chandra, “Prediction of student academic performance using machine learning algorithms,” in *Proceedings of the 13th International Conference on eLearning (eLearning-2022)*, Belgrade, Serbia, 2022, pp. 35–43.

[11] R. Liu, “Data analysis of educational evaluation using K-means clustering method,” *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–10, Jul. 2022.

[12] W. Chang, “Analysis of University Students’ Behavior Based on a Fusion K-Means Clustering Algorithm, ” NATO Adv,” *Applied Sciences*, vol. 10, no. 18, 2020.

[13] S. Souabi, A. Retbi, M. K. Idrissi, and S. Bennani, “A recommendation approach in social

learning based on K-means clustering," *In 2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pp. 1-5. IEEE, 2020.
[14] I. Pasina, G. Bayram, W. Labib, A. Abdelhadi, and M. Nurunnabi, "Clustering students into groups according to their learning style," *MethodsX*, vol. 6, pp. 2189–2197, Sep. 2019.
[15] *Learning and understanding*. Washington, D.C.: National Academies Press, 2002.
[16] A. A. Sewagegn, "Learning objective and assessment linkage: Its contribution to meaningful student learning," *Universal Journal of Educational Research*, vol. 8, no. 11, pp. 5044–5052, Oct. 2020.
[17] R. Dumbraveanu and L. Peca, "E-learning strategy in the elaboration of courses," in *Proceedings of the International Conference on Virtual Learning - VIRTUAL LEARNING - VIRTUAL REALITY (17th edition)*, 2022.
[18] J. Jovic, M. Rasoioivic Milic, S. Cvetanovic, D. Domazet, R. M. Nikolic and E. S. Vejar., "Intelligent Recommender System for Personalized Online Learning"," in *Proceedings of the 11th International Conference on eLearning (eLearning-2020)*, Belgrade, Serbia, 2020, pp. 24–25.
[19] Y. Li and H. Wu, "A clustering method based on K-means algorithm," *Physics Procedia*, vol. 25, pp. 1104–1109, 2012.
[20] G. Ivosev, L. Burton, and R. Bonner, "Dimensionality reduction and visualization in principal component analysis," *Analytical chemistry.*, vol. 80, no. 13, pp. 4933–4944, Jul. 2008.