Education and Information Technologies (2023) 28:7265–7290
https://doi.org/10.1007/s10639-022-11468-9

Post hoc identification of student groups: Combining user modeling with cluster analysis

Igor Balaban[1] · **Danijel Filipović**[1] · **Miran Zlatović**[1]

Received: 8 June 2022 / Accepted: 14 November 2022 / Published online: 30 November 2022

Abstract
This study aims to discover groups of students enrolled in the emergency remote teaching online course based on the various course-related data collected throughout the first year of COVID-19 pandemic. Research was conducted among 222 students enrolled in the course "Business Informatics" at the Faculty of Organization and Informatics of the University of Zagreb in the academic year 2020/2021. Overlays were used to model students' success on the various quizzes and exams within the course. The k-means clustering was employed to classify students into groups, based on combination of students' overlay values, frequency of accessing course lessons and the final grades. Three distinct clusters (i.e., students' groups) were discovered and explained in the given context. The identified groups of students can be used for future adaptations of the online course design in order to improve the retention and their final grades.

Keywords Emergency remote teaching · Student activity · Overlay model · Clustering

1 Introduction

The COVID-19 pandemic has accelerated the digitalization in higher education institutions (HEI) and has acted as a change driver in teaching and learning practice. Emergency remote teaching (ERT) that arose as a new phenomenon (Hodges et al., 2020) is described as an immediate change of conventional teaching practice through

✉ Miran Zlatović
miran.zlatovic@foi.unizg.hr

Igor Balaban
igor.balaban@foi.unizg.hr

Danijel Filipović
danijel.filipovic@foi.unizg.hr

[1] Faculty of Organization and Informatics, University of Zagreb, Pavlinska 2, 42 000 Varaždin, Croatia

the application of online tools. Teachers extensively relied on the institutional virtual learning platforms and videoconferencing systems which became foundational to the education experience, highlighting the transfer from onsite teaching practice into online (Bond et al., 2021; Lowenthal et al., 2020). More research is needed on the topic since the ERT presents both technological and pedagogical change. Many HEIs did not have time to properly design and adapt their courses for online mode since the shift "… has been an abrupt one due to unprecedented lockdown imposed to manage the COVID-19" (Muthuprasad et al., 2021).

Although the advanced adaptive online education techniques and systems could help to improve student retention rates (Smaili et al., 2021), the hard reality of many institutions which were forced to shift to emergency online teaching is that they do not have the access to or cannot use the adaptive online education platforms. In situations where the curricula were transformed to online form within the traditional non-adaptive paradigm (i.e., basic Learning Management Systems (LMS) usage), it is not possible to continuously adapt the learning process to the individual needs of students. It seems to be a reasonable alternative that educators can use the data from the LMS systems collected during the initial years of crisis to perform post hoc user analysis and identify different groups of students in their courses. Consequently, they can refine course content and structure to offer adaptations in future in order to improve students' retention rate.

The various forms of resource usage analysis and user modeling have been used for more than a decade as a foundation for redesigning online courses (i.e., in Camacho et al., 2009; Mazza & Botturi, 2007). Learning Management Systems (LMS) can thus present a valuable tool to help teachers to analyze student work and activities in online courses. Many resources describe how logs from such systems can be used to research the relation between the time spent online and students' final grades (Ryabov, 2012), to research the impact of online learning activities on learning outcomes in blended courses (Nguyen, 2017) or to explore the relationship between the time students spent on the course website and the assessment performance (Korkofingas & Macri, 2013).

This research aims to classify students enrolled in the online course during ERT period, based on the post hoc analysis of the available data about the students' activities throughout the entire semester and on their mastery of knowledge domains. To achieve this, the final course grades were correlated with the results of two types of activities: (i) the knowledge displayed on various online tests during the semester (many formative and summative online tests within the course - multiple flash exams and self-assessments, two midterm and the final exam) and (ii) the frequency of reading the provided lessons (i.e. students' activity). Such identification of groups of students should allow teachers to adapt the structure of teaching and learning activities in the course, aiming to improve the students' retention rate and their final grades. In this paper, we define the term 'retention' as the number of students who continue their studies by re-enrolling from one academic year to the next, i.e. we are referring to the students' retention rate or academic retention. The term 'retention' should therefore not be interpreted as knowledge retention or memory retention.

Considerations towards improving academic retention are valid, especially at the beginning of a study programme. Some studies have shown that students at

the beginning of their academic career have problems with assessing the workload brought by certain courses, and that failure/not finding the right way at the beginning of their studies can negatively affect the students' retention rate. According to Otrel-Cass et al. (2009), "… findings suggest that during their first year science students need to be reassured that they are valued, and that their education is taken very seriously by the institution and their lecturers. Student commentary suggests this can be achieved by personalizing lectures, ensuring personal contact with lecturers and monitoring how students are coping with the challenges and stresses that affect workload issues and subsequently their academic progress.". This finding is relevant for the course which is subject of this research, since our course is one of the foundational courses within the curriculum and it takes place in the first semester of a three-year undergraduate professional study program.

2 Theoretical background

According to Kebritchi et al. (2017), it is a great challenge to develop online courses which not only cover the curricular aspect, but also succeed in engaging the students and preparing the instructors for transition to online teaching and learning. Bignoux and Sund (2018) have identified major differences between the online environments and traditional classrooms regarding the student's satisfaction, motivation and interaction. Other studies have shown that the effectiveness of online classes depends not only on the use of advanced technology, but also on the quality of course content structure, instructors' preparedness and the quality of instruction (Sun & Chen, 2016; Gilbert, 2015).

2.1 Using LMS data to analyze students' activity

LMS systems record vast amounts of logs containing various details about students' activities during course duration (i.e., what activities students have accessed, when, for how long, etc.). Teachers should be, therefore, empowered to collect, analyze and interpret the data collected in the LMS with learning analytics (LA) tools to improve students' learning (see for example Rapanta et al., 2021). The necessity of LA usage is further highlighted by Ferguson (2012) who noted the importance of using the data about students and their contexts to understand and optimize the learning environments, as well as the learning process. For example, Ryabov (2012) reports that the overall logged time within LMS (i.e., time spent online) had a positive influence on the final grade. Literature review by Nguyen (2017) reports positive correlations between readings of various contents (i.e. pages viewed, read discussions, discussion posts made) and the learning outcomes. Wei et al. (2015) impact analysis of the activity within LMS systems on the academic performance has shown that the results of various online assignments (including online tests/exams), as well as the overall access time (i.e., overall number of logins, number of posts, time spent to read various documents) had significant effect on the learning performance.

2.2 Approaches to identification of students' groups

As noted in the previous section, one of the important research streams in the analysis of student data in LMS is the user modeling, especially in the context of adaptive education to explain the student groups and thus to better adapt the course learning design (i.e., in Corrin et al., 2017). User modeling is defined as a process of obtaining a user model as "… a source of information, which contains assumptions about those aspects of a user that might be relevant for behaviour of information adaptation." (Schreck, 2003) In education, such approach includes overlay student models, stereotypes, perturbation models, machine learning techniques, constraint-based models, fuzzy models, Bayesian networks and ontology-based models. Hybrid student models in which researchers combine various modeling techniques have also been recorded. However, not every approach is suited to model every characteristic of a student. Chrysafiadi and Virvou (2013) review reveals that a student's mastery of knowledge is predominantly modeled with overlays (20% of all research) and stereotypes (14.44% of all research).

The overlay model assumes that the student has incomplete but correct knowledge of the domain, i.e., the model of student's knowledge is a subset of the domain model. The domain model represents the expert-level knowledge of the domain and it is compared with the model of student's knowledge. The difference between them is believed to originate from the student's lack of skills and knowledge. Therefore, the main objective of the instruction process is to eliminate these differences as much as possible. Essentially, the domain must be decomposed into groups of smaller, interconnected elements (knowledge topics) and, therefore, the individual student's overlay model consists of a set of masteries (student's recorded levels of knowledge) of those elements. Masteries can be expressed as Boolean value (true/false – i.e., student does or does not possess the knowledge about an element) or more refined qualitative (i.e., good/average/scales) or quantitative measures (i.e., probability level that student has the knowledge). (Martins et al., 2008; Nguyen & Do, 2008; Brusilovsky & Millán, 2007; Bontcheva & Wilks, 2005).

More modern approaches also include various data mining and classification techniques (i.e., decision trees) and clustering techniques (Francis & Babu, 2019) that enable real-time grouping.

3 Research objectives and methodology

This research aims to discover possible groups of students enrolled in an online course[1] based on post hoc activity data collected from LMS logs (frequency of accessing course lessons, quiz scores) and students' final grades. Several studies have shown that the post hoc analysis of LMS data combined with clustering techniques can be used to derive student groups, i.e., by analyzing self-assessment scores

[1] Course is held fully online as part of the emergency remote teaching process. Under normal circumstances it would be held as a blended course.

(Watson et al., 2017) or by analyzing students' activities including login records and content reads (Tseng et al., 2016).

3.1 Research objectives

In this research the students' knowledge and the overlay model are used to calculate students' masteries post hoc, when all the lectures in a course are finished. Calculation of masteries is based on the students' scores obtained within a number of course online quizzes (including informal self-assessments and formal midterms/exams). Clustering is used to classify our students into typical groups based on their masteries from overlay model, activity logs data and final exam grades.

The research was conducted during an emergency lockdown period, when many pedagogical and technological aspects of teaching were pushed forward by institutions (e.g., asynchronous and synchronous teaching and learning, prescribed LMS and video streaming platforms, etc.). Although the transition to online teaching was mostly ad-hoc for many courses, LMS platforms enabled systematic recording of the activities of the learners.

With respect to the research aim and the above-mentioned context, the following main research question has been formulated:

> What types of student groups can be detected based on calculated students' masteries post hoc, activity logs and final exam grades?

For that purpose, several associated research objectives were set:

1. To calculate students' masteries post hoc using an overlay model.
2. To classify students into groups based on the overlay model, activity logs and final grades using clustering.
3. To describe these student groups, with implications for improving the course structure and activities.

3.2 Research process

The research process has been divided into three phases:

1. Course preparation.
2. Data collection and preprocessing.
3. Data analysis.

In the first phase, the online course was prepared according to the principles of programmed learning in LMS Moodle. The variety of online resources and knowledge assessments were implemented along with conditional activities to create a clear learning path for students. More information about the used programmed learning principles, conditional activities and the course in general is given in the subsequent Section 4.

Within the *data collection and preprocessing* phase, raw data about various activities of students have been gathered from different sources. The dataset consisted of three parts: (1) the overlay model of students' domain knowledge, (2) students' frequency of viewing course material – students' activities, and (3) students' final course grades. A special database was designed to accumulate the collected data and provide a more structured and organized data source to generate the desired dataset. Python programming language, SQLAlchemy database toolkit and a SQLite database were used to implement and work with the database.

The data for the overlay model was collected from the Moodle LMS by exporting the questions and students' responses from all quiz activities (midterms, flash-tests, self-assessment quizzes, etc.). Students' activities were collected from course activity logs on Moodle LMS. Students' final course grades were collected from two data sources: (1) from our own automatic grading sheet implemented in Microsoft Excel, and (2) from the national HE information system (ISVU) exported as an Excel sheet. Grades from (1) were the grades students achieved during the semester. Grades from (2) were the grades students achieved during multiple post-semester examinations.

In the end, a pivot table was created as the main dataset for analysis, where the columns Student and Center were used as indices, column Property name was used to generate the column headers, and Property value was used to display the pivot table values. The structure was as follows:

- *Student* − student's full name.
- *Center* − the center (cities where our course is taught, more details in subsequent Section 4) student belonged to.
- Student's self-assessment scores for the knowledge domains (more details on knowledge domains can also be found in subsequent Section 4):
  - *Information systems* (IS)
  - *Memory unit* (MU)
  - *Basic computer principles* (BCP)
  - *Computer software* (CS)
  - *Information system security* (ISS)
  - *Central unit of a computer* (CUC)
  - *Input/Output unit* (IOU)
- *Lessons 1.1–1.5* (L1.1-L1.5), *Lessons 2.1–2.18* (L2.1-L2.18), *Lessons 3.1–3.4* (L3.1-L3.4), *Lessons 5.1–5.2* (L5.1-L5.2) − student's activity in lessons (reading materials for all knowledge domains, more details on lessons in subsequent Section 4).
- *Grade* − student's final grade.

The presented dataset was used to:

1. Perform the correlation analysis between the main components within the dataset (i.e., overlay model values, students' activities and final course grades).

2. Analyze the entire dataset by means of clustering, using the Silhouette method to determine the optimal number of clusters and k-means clustering algorithm to form the actual clusters.
3. Apply descriptive statistics to determine the meaning of each cluster, i.e., what kind of student group each cluster may represent.

Having a full dataset available, the first step within the *data analysis* phase was to analyze the correlations between the main components of the data set (i.e., overlay model values, students' activities and final course grades). Afterwards, the entire dataset was analyzed by means of clustering, using the Silhouette method to determine the optimal number of clusters and k-means clustering algorithm to form the actual clusters. Having the clusters formed, descriptive statistics was used to determine the meaning of each cluster, i.e., what kind of student group each cluster may represent.

In order to assess the homogeneity of grade distribution across academic years and centers, a log-linear model was analyzed using generalized linear regression with Poisson distribution and the natural logarithm link function.

4 Course description

The course "Business Informatics" is taught in the first semester of a three-year undergraduate professional study program at the Faculty of Organization and Informatics of the University of Zagreb. The study program is transdisciplinary, including the fields of informatics and economics. The syllabus of "Business Informatics" covers several major units of content – overview of the information systems and their business applications, deeper insight into computer hardware and software (basic elements of information systems) and the basics of information systems' security.

Under normal circumstances "Business Informatics" is a blended course held at four different locations in Croatia - at the main Faculty location in Varaždin (hereafter, Main Center - MC) and at three additional dislocated study centers in other towns in Croatia (hereafter, Dislocated centers - DCs). However, during the past five years the course was also piloted as an online course in DCs and with that respect, the materials as well as the methods were prepared for online teaching and learning. Therefore, during the COVID-19, the course "Business Informatics" was carried out as a fully online course, for MC and DCs according to the methodology used in previous years for DCs. Onsite classes were replaced by synchronous[2] and asynchronous online activities. All asynchronous activities were offered as digital contents within LMS (reading materials, quizzes, self-assessments), supplementing the synchronously delivered online lectures. Students could have worked at those activities at their own pace, without strictly enforced schedules or deadlines.

[2] Classes were live-streamed using video conference tools, according to the formal schedule.

During the academic year 2020/2021, the course was enrolled by 222 students in total, out of which 101 were enrolled in MC and 121 in DCs.

4.1 Course design

Synchronous and asynchronous parts of the course were administered in LMS Moodle and its learning design aspects were organized according to the basic principles of programmed learning paths, as summarized by Seel (2012):

- learning contents are broken down into smaller pieces of content, which are immediately followed by one or more comprehension questions.
- student receives immediate feedback about the correctness of the answers.
- if the answers are correct, students can proceed to the next piece of content.
- if the answers are not correct, students are required to revise the content and answer the questions again.

The asynchronous materials for online classes were prepared for every part of the syllabus. Students are supposed to use the asynchronous programmed learning path of the course activities to relearn already taught course topics at their own pace (i.e., topics covered during synchronous online lectures), or to use it as a primary source of learning about topics which were planned for asynchronous self-studying.

The technical part of the programmed learning path was facilitated by the built-in Moodle features, primarily by Lesson activities. Lessons in Moodle allow combining the content pages (text + multimedia) and questions with feedback into a single unit which was found to be ideal for creating basic building blocks of the programmed instruction within the "Business Informatics" course.

According to Britain (2004), the learning workflow originating from a learning design enables the teachers to create more structured teaching activities, thus leading to more effective learning. Both Britain (2004) and Dohn (2010) stress that the inclusion of learning designs into a course puts the learning activities in focus and provides a framework for deep reflection during the course design process.

Every major unit of content was organized as the domain of knowledge in the following way (a programmed learning path):

- Unit topics were broken down into a sequence of lessons.
- Each lesson consisted of several pages and finished with several questions that all must be answered correctly. The immediate feedback was given after each question and if the answer was not correct, students could retry the question or return to the content pages within the lesson.
- Access to the following lesson was allowed only when all the questions of a previous lesson had been answered correctly (i.e., 100% completed previous lesson).
- Every major unit (domain of knowledge) ended with a final self-assessment quiz which covered all the lessons within the unit and assessed the domain knowledge. Access to the self-assessment was possible only when a student

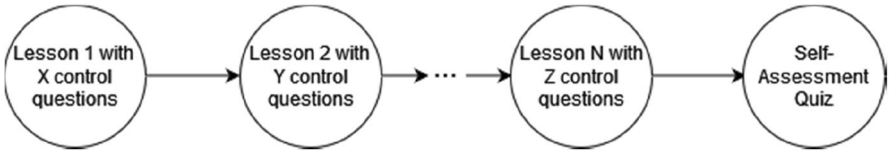

Fig. 1 Lessons flow within major units of content, followed by final self-assessment

achieved 100% completion rate within all lessons in that unit. The number of attempts was unlimited.

Figure 1 additionally shows the sequential flow of the conditional activities within any major unit/domain knowledge of content. Conditional activities (technical features of the Moodle LMS) were used to control students' flow through the elements of each domain of knowledge. Access to most of the resources in Moodle (lessons and quizzes included) can be based on students' individual results achieved on any other LMS resource within the course (i.e., another lesson, quiz, assignment, forum participation, etc.). These conditional activities have been used to prevent students from accessing further elements within the domain of knowledge, until they had shown the required mastery of the previous elements. The precondition for accessing the following lesson was achieving the required percentage of correct answers to the control questions at the end of a previous lesson, while the access to the final self-assessment quiz was conditioned by achieving required success in all lessons within the domain.

All major units of knowledge, organized as shown in Fig. 1, were available in LMS to provide the asynchronous portion of the classes. Students from all centers (MC and DCs) had equal access options to them.

For completing the final self-assessment for a domain knowledge with at least 75% success, a student has been awarded a badge, which proved that a certain level of mastery within that unit was achieved. Ideally, each student would earn a badge for every major unit in course.

In total, 7 major units or domains of knowledge are covered in the course, as shown in Fig. 2: Information Systems, Information Systems Security, Basic Computer Principles, Central Unit of a Computer, Memory Unit, Input/Output Unit, and Computer Software. The mappings between domains of knowledge and lessons are also highlighted.

Named circles at the beginnings of the two chains in Fig. 2 represent domains of knowledge which are mutually independent, and students may choose to begin the asynchronous portion of studying from either of these. Subsequent circles within the chains represent domains which depend upon previous domains and should be studied after the required level of mastery in previous domains is acquired. For example, to advance to the "Memory Unit", students must first study "Basic Computer Principles" and then "Central Unit of a Computer".

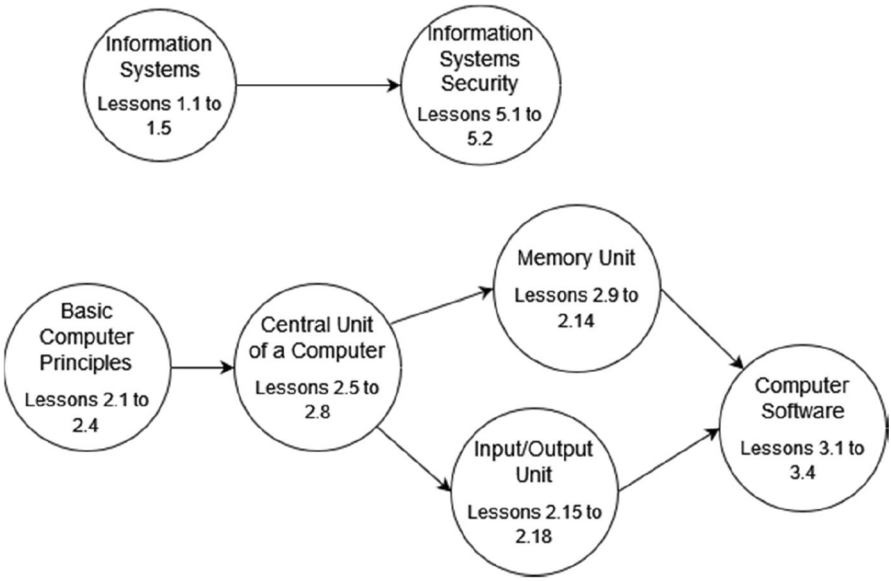

Fig. 2 Dependencies between the main domains of knowledge (units of learning contents)

4.2 Formal knowledge assessments

Besides the mentioned self-assessments for domain knowledge, the formal online midterms and the final exam had to be taken. For the midterms (during semester) and the final exam (after the semester has ended) the formal online tests were prepared within the same Moodle course (new tests for each midterm/exam period). The grades from all those formal tests were indirectly included in research as part of students' final grades.

5 Results

Dataset analysis started with the calculation of correlation coefficients to understand the relationships between the main components of the dataset (overlay model, students' activity, and final grade), i.e., how much effect do variables from one component have on the variables in other components. Next, the cluster analysis was conducted on the dataset.

Before cluster analysis, the Silhouette method was employed to determine the optimal number of clusters for the clustering process. Secondly, after dataset clustering, a total number of students per cluster and distribution of final grades per cluster were counted as a form of introductory insight into the clustering result. Finally, descriptive statistics were applied to analyze the characteristics of individual clusters, i.e., to determine what kind of student group each cluster represents.

5.1 Correlation analysis between the main components of the dataset

The correlations have been analyzed in programming language Python using the pandas data analysis library and visualized with the seaborn data visualization library as a heatmap.

Firstly, the Shapiro-Wilk test was applied on all variables to check the dataset for normality (Yap & Sim, 2011). Since none of the variables were normally distributed, the Spearman correlation was used due to the several characteristics that are of interest for this analysis (Schober et al., 2018):

1. it is useful for non-normally distributed data.
2. it uses rank of values of the variables with calculated correlations, instead of the actual values and.
3. can be used for ordinal data.

For (1), Shapiro-Wilk test was already used to determine that none of the variables have normally distributed data. For (2) and (3), since the values of all the variables (except for Grade) are in range between 0 and 1, the actual values are converted into ordinal ranks. Table 1 shows the conversion information, as well as the description of what a rank means for the type of variable. Variable Grade is already considered to have ranked values.

Figure 3 shows the Spearman correlation of the dataset as a heatmap. It can be noted that there is a segregation between the subsets of data. While the variables from the overlay model (self-assessments IS, BCP, CUC, MU, IOU, CS and ISS) clearly have a medium-to-large effect on each other (from 0.5 to 0.8), they mostly display a small effect (0.2 to 0.4) on the variables that form the student activity (reading of lessons, L1.1 until L5.2). Likewise, the variables that form student activity have a larger effect on each other (from 0.4 to 0.8) and mostly a small effect on the variables of the overlay model (from 0.2 to 0.4). It can be concluded that, for the majority of the students, accessing lessons does not contribute greatly to their understanding of the domain knowledge.

Furthermore, it is important to note the relationship between the overlay model, the student activity variables and the Grade variable. While the overlay model variables have a medium-to-large effect on the grade (from 0.4 to 0.6), the effect of the student activity on the grade is low (from 0 to 0.3). Based on these findings it can be

Table 1 Conversion information for the Spearman correlation

Actual value	Ranked value	Description for the overlay model	Description for the students' activity
0.0–0.2	1	Lowest knowledge of the domain	Lowest activity on lesson
0.2–0.4	2	Low knowledge of the domain	Low activity on lesson
0.4–0.6	3	Moderate knowledge of the domain	Moderate activity on lesson
0.6–0.8	4	High knowledge of the domain	High activity on lesson
0.8–1.0	5	Highest knowledge of the domain	Highest activity on lesson

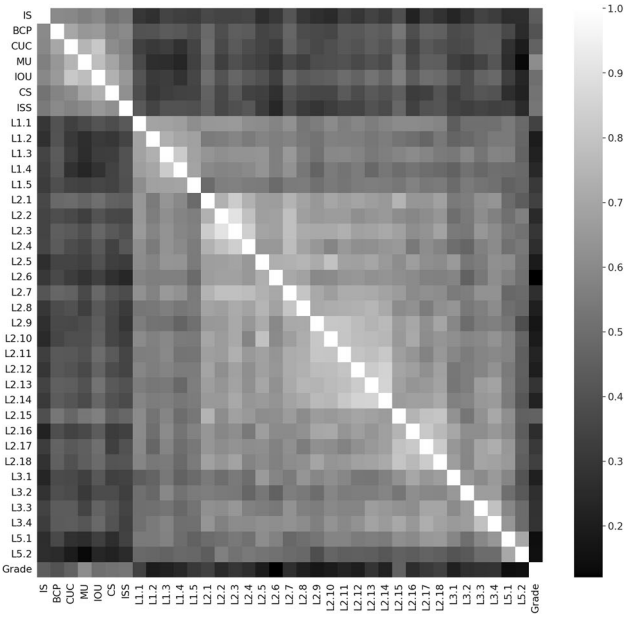

Fig. 3 Spearman correlation of the dataset

concluded that the level of knowledge affects the student grade but the frequency of reading the lessons doesn't significantly affect the increase in their final grade.

5.2 Cluster analysis of the dataset

The cluster analysis was performed by application of the k-means clustering algorithm on the dataset followed by the content analysis of the clusters. The k-means algorithm is a non-hierarchical clustering algorithm that can be used for various tasks and analyses in various data-related fields. As reviewed by Dutt et al. (2017) this algorithm is frequently used in the field of educational data mining.

However, before application of the k-means algorithm, the Silhouette method was used to determine the optimal number of clusters (i.e., in Rousseeuw, 1987; Chiang & Mirkin, 2010; De Amorim & Hennig, 2015). For determining the optimal number of clusters and clustering, the data in the student activity subset of variables have been scaled due to the extremely high or low values for some students.

The dataset has been clustered nine times, each time with a different number of clusters: starting with two and ending with ten clusters. Then, using the scikit learn library for the Python programming language, specifically the metrics module, the Average Silhouette Width (ASW) was calculated for each clustering process. Finally, the ASW scores were plotted to visually determine the optimal number of clusters. The Silhouette method plot shown in Fig. 4 was used to detect the optimal number of clusters. The highest score of 0.44 was given to the clustering with three clusters, which was selected as the optimal number of clusters for clustering analysis. The three clusters were labeled C1, C2, and C3 and were further analyzed.

Fig. 4 Silhouette method plot

Table 2 Final grades for each cluster

Grade	Clusters			Total
	C1	C2	C3	
1	1	125	22	148
2	37	0	7	44
3	23	0	4	27
4	2	0	1	3
5	0	0	0	0
No. of students	63	125	34	222
No. of dropouts[a]	2	21	3	26

[a]Students who left the study program within 9 months after the course ended

5.2.1 Final grades per cluster

The frequency of each individual grade was counted for each cluster and displayed in Table 2. Grades[3] range between 1, representing the lowest grade (meaning students failed the course) and 5 representing the highest.

Cluster C1 consists mostly of students with the passing grades except for one student that failed the course. In comparison, C2 includes only the students that have failed the course. Finally, C3 contains a smaller number of both, students that passed and the students that failed the course. Regarding the student dropouts, cluster C2 includes the highest number of dropouts (21 dropouts which is 80.77% of the total number of dropouts in the course). The characteristics of C3 will be elaborated after further analysis in the following sections.

[3] In Croatia the following official grade scale applies to elementary school, high school and university students: 1-insufficient/failed, 2-sufficient, 3-good, 4-very good, 5-excellent.

5.2.2 Standard deviations per cluster

In this section, the standard deviation for each cluster is calculated and displayed within the charts in Fig. 5. Figure 5a shows the standard deviations of the overlay model part of the dataset and Fig. 5b shows the standard deviation of the students' activity. In these cases, the values for the overlay model and the values of the students' activities are scaled to range between 0 and 1, so the calculated standard deviations are in the same range.

Cluster C1 consists mostly of students earning their passing grades (grades 2, 3, and 4), with one exception, and with the standard deviations fluctuating around 0.2 in the overlay model. The standard deviations of students' activities fluctuate around the 0.1 value which could point to the fact that the clustering algorithm was more dependent on students' activities than domain knowledge. The deviations not being closer to zero could also be attributed to not all students in the cluster making the same effort, but still earning the passing grade. A student that earned 2 as the final grade probably didn't fully understand the domain knowledge and/or wasn't very active in studying the course materials. On the contrary, a student that earned 4 as a final grade probably understood the domain knowledge to a greater degree and/or was actively studying through the lessons.

In the case of cluster C2, the standard deviation for the overlay model for the first 4 knowledge domains is above 0.2 (with one domain above 0.3), while for the

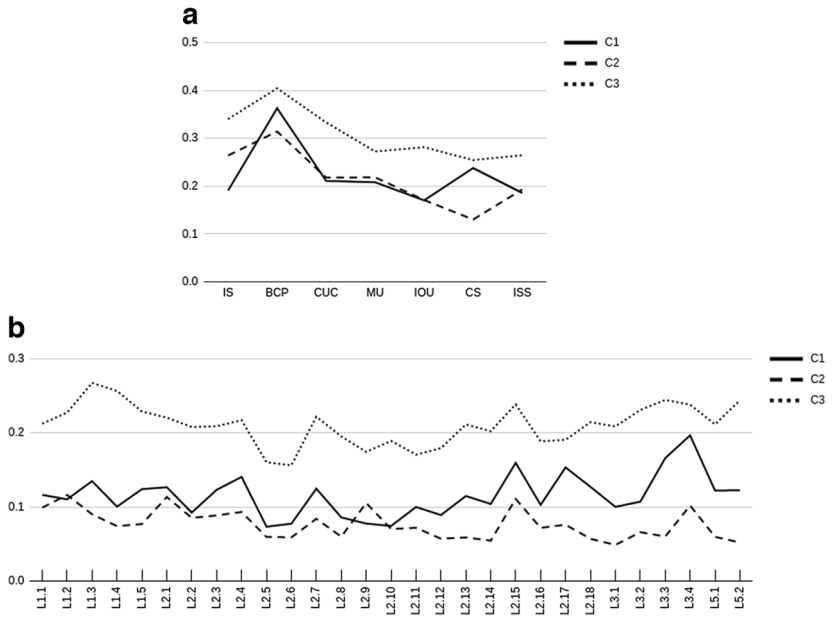

Fig. 5 **a** Standard deviation for the overlay model for each cluster. **b** Standard deviation for students' activities for each cluster

remaining 3 domains the values are around 0.2 or lower. This is the cluster that contains only the students who failed the course. Low standard deviations in the students' activities (mostly below 0.1) suggest that the clustering algorithm depended mostly on students' lower activities and failing grade.

Standard deviations for the cluster C3 are higher than in the other two clusters which can be explained by more varied final grades and the self-assessment results of these students. Additional analysis will show that this cluster also contains the most active students, which in conjunction with varying final grades (ranging from 1 to 4) could also explain the higher standard deviation in the overlay model.

5.2.3 Mean and median comparison per cluster

In this section, the mean and median values were calculated for each attribute of each cluster and displayed in two charts: (1) for the overlay model part of the data (see Fig. 6), and (2) for the students' activities part (see Fig. 7). The levels of domain knowledge and the levels of activity are presented using the ranks from Table 1.

In the overlay model of the cluster C1, the means and medians show moderate levels of knowledge (0.4 to 0.6 range) for all domains except for one, having low level of knowledge (0.2 to 0.4 range). The students' activities in this cluster falls into the lowest activities range (0.0 to 0.2) for all lessons but one. However, in comparison with the C2, C1 is still a more active cluster.

For the cluster C2, means and medians of the overlay model show the lowest level of knowledge for all domains (all below 0.1). Medians for all domains are zero, indicating that at least half of all the students in this cluster have the lowest level of knowledge of the domains (0.0 to 0.2 range). This cluster also shows the lowest students' activities levels, consistently being below 0.1. Starting with the lesson L1.3 median values are zero, indicating that from that lesson onwards at least half the students in the cluster C2 stopped accessing course materials.

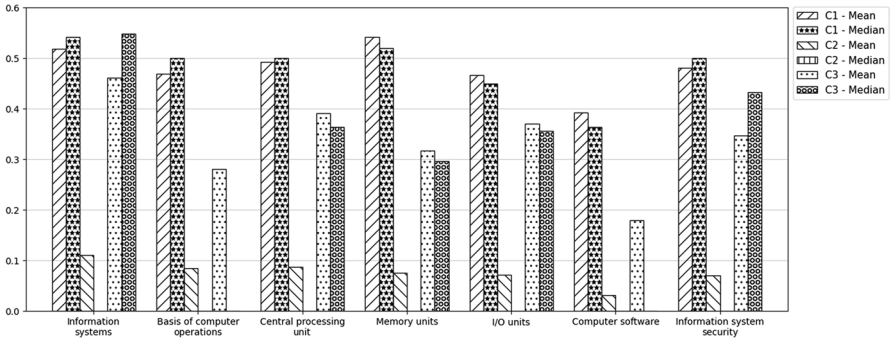

Fig. 6 Mean and median values of the overlay model for all three clusters

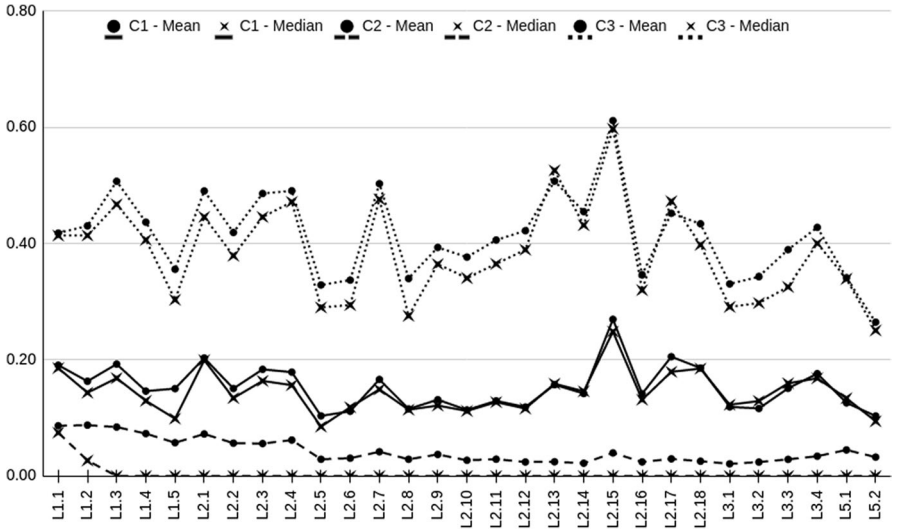

Fig. 7 Mean and median values of students' activities for all three clusters

There are notable inconsistencies in mean and median values for the C3 cluster's overlay model. Means range from the lowest levels of the domain knowledge (slightly below 0.2) to the moderate levels (0.4 to 0.6). Median values are zero for two domains (see Fig. 6), indicating that at least half of the students in C3 have shown the lowest level of knowledge of these domains (0.0 to 0.2 range). Regarding the students' activities in C3 cluster, both means and medians are relatively equally distributed between low (0.2 to 0.4) and moderate (0.4 to 0.6) levels. Figure 7 clearly indicates that the cluster C3 is the most active of all three clusters. These data support the initial assumption that the cluster C3 contains mostly the students who try to compensate for the difficulties in understanding the domain knowledge by the increased activity (i.e. more frequent reading of lessons).

5.3 Additional analyses of the clusters

Additional analyses were made on the clustered dataset to examine the per-cluster distribution of students from all centers (MC and DCs) and the frequency of taking self-assessment quizzes by clusters and knowledge domains.

5.3.1 Distribution of students from study centers by clusters

Table 3 shows the distribution of students from the study centers by clusters.

From the perspective of the clusters, 88.89% of the students (56 out of 63) in the cluster C1 are enrolled in the MC. Most of the students in the cluster C2 (75.2%, 94 out of 125) belong to the dislocated centers (DCs). The most equal per-center distribution is noted for cluster C3 - approx. 42% of students (14 out of 34) belong to MC and approx. 58% (20 out of 34) belong to DCs.

Table 3 Distribution of students from each study center by clusters

Cluster	MC	DC	Total (clusters)
C1	56	7	63
C2	31	94	125
C3	14	20	34
Total (centers)	101	121	222

The analysis from the perspective of the centers gives better insights, particularly when it comes to identifying the problematic centers. Students from the DCs are placed predominantly in the clusters C2 (94 out of 121 students, 78%) and C3 (20 out of 121 students, 16%). Students from the MC are mostly distributed in clusters C1 and C2, with cluster C1 having the dominant position (56 out of 101 students, 55% in C1 vs. 31 out of 101 students, 31% in C2).

According to this data, it can be concluded that the DCs are more problematic centers, having the majority of students (78%) in the cluster C2, representing the students that failed the course. The MC has approximately 55%/45% split between the successful students (in C1) and the students that either failed the course (in C2) or had difficulties in their studies (in C3), which could indicate there is still room for improvement in course design and more active monitoring of such students.

5.3.2 Frequencies of taking self-assessment quizzes by clusters

Figure 8 shows the frequencies of taking self-assessments, which are sorted in order defined by the course curriculum. Only 137 out of 222 students in the dataset have been taking the self-assessment quizzes. As mentioned in the Subsection 4.1, the number of attempts to solve the self-assessments was not limited.

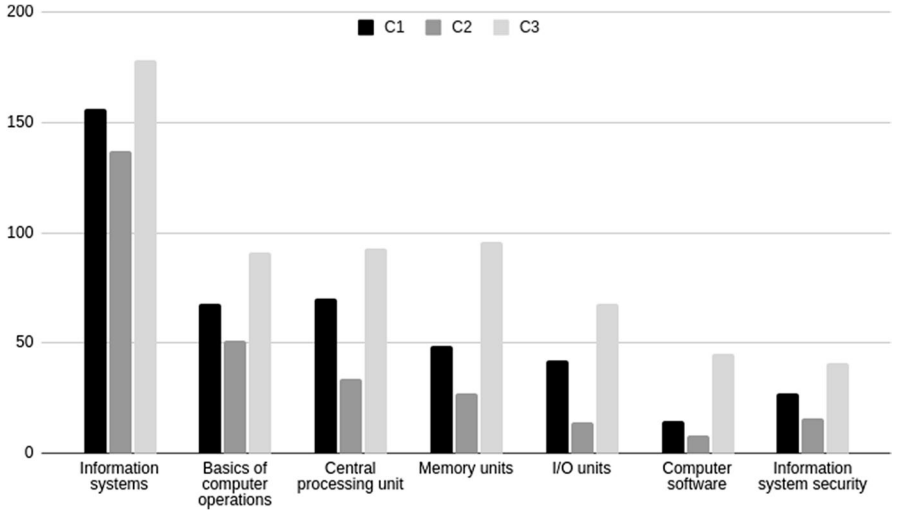

Fig. 8 Frequency of taking self-assessment quizzes by clusters

The Information systems self-assessment quiz (the first self-assessment quiz chronologically available in the course) shows the highest frequency of attempts. Also, it is the only quiz with more than 100 attempts of solving in all clusters. This could be attributed to different reasons: students starting the course with high hopes, interest and/or confidence, students having difficulties with understanding the domain knowledge and taking the quiz countlessly until they get the answers correct, etc. A significant decrease in number of attempts for the next 3 self-assessment quizzes and even greater decrease for the last 3 was recorded. This could be attributed to either students losing interest in the self-assessment quizzes (especially in the cluster C2) or to students having less difficulties with understanding the topics.

From the perspective of the clusters, we can see that the cluster C1 (which includes the students that, with one exception, have all received passing grades and are considered to have no difficulties with understanding the domain knowledge) has the second highest frequency of taking self-assessment quizzes. The lowest frequencies of taking the self-assessments are noted in the cluster C2 (students that failed the course and had low activity in accessing course materials). The cluster C3, with the most active students having various but mostly failing grades, also has the highest frequencies in taking the self-assessment quizzes.

These frequencies reinforce the idea that the cluster C3 includes the students that have difficulties with understanding domain knowledge and spend more time studying in order to compensate.

6 Discussion

The dataset that was used in the analysis was constructed from students' overlay model, their frequency of reading materials in Moodle LMS and the earned final grade. Next, a clustering technique was used to generate clusters which were further analyzed to determine student groups. After determining the optimal number of clusters with the Silhouette method, three clusters were formed from the dataset and identified as the following student groups:

- Cluster C1 contains 63 students, out of which 62 have passed the course. Cluster standard deviations show a low variance in data when it comes to students' activities within the LMS. The mean and median values for the students' activities report they had a low to medium level of activity within the course. Overall, C1 seems to contain somewhat active students with better understanding of the domain knowledge than the students in other clusters.
- Cluster C2 contains 125 students, all of which have failed the course. Standard deviations also show a low variance in data regarding students' activities in LMS. Cluster's mean and median values show that at least half of the students had the lowest level of domain knowledge, as well as not being active in the course. Overall, C2 seems to contain the students that have failed the course, had low or no understanding of the domain, and were either slightly active or not active at all.
- Cluster C3 was the most intriguing for further analysis. It contains 34 students, out of which 12 have passed and 22 have failed the course. Its stand-

ard deviations show a higher variance in comparison to C1 and C2. Results show that at least half of the students had the lowest level of knowledge of two domains (*Basics of computer operations* and *Computer software*), a low level of knowledge for three domains (*Central processing, Memory* and *I/O units*) and moderate level of knowledge for two domains (*Informations systems* and *Information systems security*). The mean and median values place them in the moderate level of activity. Overall, C3 seems to include students that were very active on the course (this seems to be their primary characteristic), but a higher percentage of them failing the course, and having difficulties understanding most of the domains or having difficulties with studying. These characteristics could mean that the C3 consists of students that are struggling with the course. Regarding the grade, they either barely got the passing grade or have the great potential of getting it.

When comparing activities of students in cluster C3 with the activities of students in other two clusters, we can observe the following differences, which support the conclusion about students in C3 having difficulties with understanding knowledge domains or with the studying process:

- Figure 7 shows that the students from C3 had the highest number of recorded access to each of the 29 lessons in LMS. When compared with C1 (students with mostly positive grades, see Table 2), the cluster C3 containing almost 50% less students than C1 (34 from C3 vs. 64 from C1, see Table 3) has achieved ranks which are at least twice as high as ranks from C1. This indicates that the students from C3 have been reading all the lessons more often than students from C1 and definitely more often than students from C2 (inactive and failing students, having median activity rank equal to 0 for 27 of the 29 lessons).
- The frequencies of taking self-assessment quizzes at the end of domains (Fig. 8) show that the students from C3 have generated the highest number of attempts to solve each of the 7 self-assessments. For the first quiz (on Information systems) it is evident that 34 students from C3 have accumulated approx. 175 attempts, averaging at approx. 5.1 attempts per student, while 64 students from C1 have accumulated approx. 150 attempts, averaging at approx. 2.3 attempts per student. Similar pattern can be observed for the other 6 quizzes. We can interpret more frequent attempts at solving these quizzes in cluster C3 as an indicator of insecurity in their knowledge, i.e. needing more attempts to achieve the positive effects of repeated quizzing on long-term retention of knowledge (Larsen et al., 2015). When these frequency-based findings for C3 are combined with the scores achieved in those quizzes (see Fig. 6), we can see that the scores of C3 are consistently one rank lower than C1 scores for 6 out of 7 quizzes. This supports the conclusion that students in C3 have difficulties in understanding most of the topics in the course.

One of the future improvements regarding the course design, especially in the context of the cluster C3 including active students who struggle with understanding would be to embed the instant feedback within online quizzes, as suggested

by Jia and Zhang (2019) in order to help students with poor academic performance in the final exam.

In order to put the results of clustering in a wider context, the distribution of grades by the center and academic year was analyzed (Fig. 9). The course takes place in the winter semester, thus the academic year 2020/21 was the first ERT year. There was variation in grade distribution among the three academic years, but the ERT year was within the pre-COVID variation in grade distribution. There was also a consistent difference in grade distribution between the centers, with more students failing in the dislocated centers. This difference was more pronounced in the ERT year. Analysis of the log-linear model with student numbers as the dependent variable, and grade, year, and center as independent variables, showed that three-way interaction was statistically significant (likelihood ratio $\chi^2 = 13.74, df = 4, p = 0.0082$). Specifically, in 2020/21 the main center had more students with grades 2 ($\beta = 1.526, z = 2.862, p = 0.0042$), and 3–5 ($\beta = 1.408, z = 2.019, p = 0.044$) in comparison to the baseline year 2018/19, dislocated center, and grade 1, after accounting for independent effects of grade, year, center, and their pairwise interactions. Therefore, it is evident that the number of failing final grades was high even before COVID-19 (esp. in DCs) and that it was not a reflection of the ERT and the COVID-19 context. It is also evident that

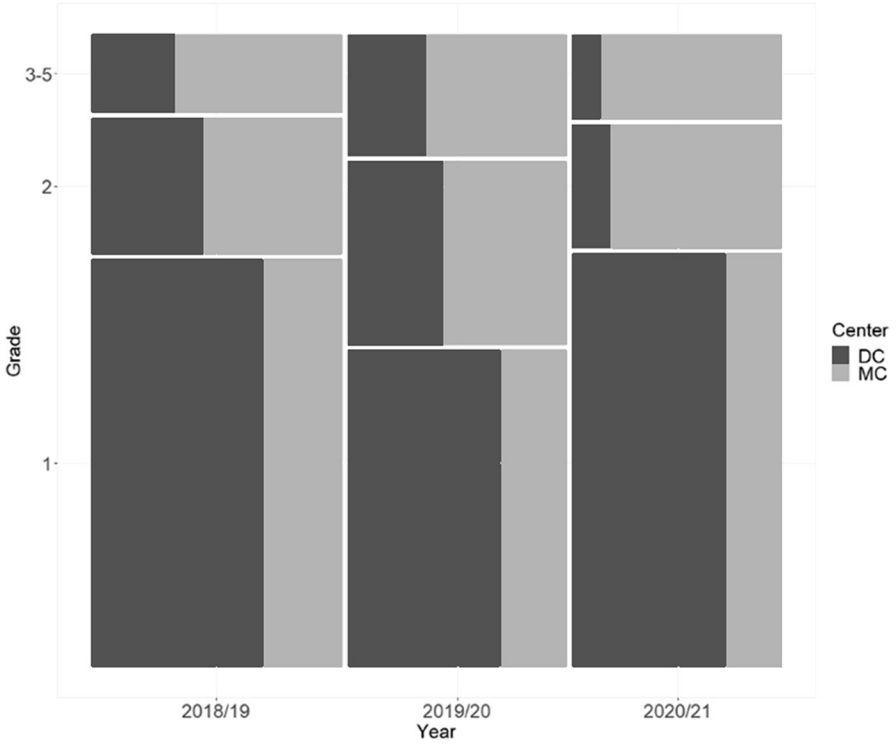

Fig. 9 Mosaic plot of the number of students by grade, academic year, and center (MC = main center, DC = dislocated center)

students from the MC have consistently shown better results than students from the DCs. The locality can also be excluded as one of the factors for allocating students from DCs primarily to the cluster C2 as students in that cluster come from all three DCs (being geographically quite distant too), as well as from the MC.

The high number of failing students within the cluster C2, which mostly includes the students from DCs, may also be explained by the insufficient individual communication with the teachers. The students from DCs mostly use asynchronous materials since the majority of them are part-time students attending or participating in the synchronous activities very rarely. Furthermore, most of them haven't used the available means of consultations with the teachers (e-mail, online synchronous consultations, and forums available in LMS). Rienties and Toetenel (2016) state that the time students spend on communication activities is one of the major predictors for academic retention suggesting that the learning outcomes should be aligned with well-designed communication activities. Therefore, for the students in the cluster C2, as well as for the failing students in the cluster C3, a series of consultations should be organized with the teacher throughout the semester to address their issues at the individual level.

Although the engagement level (i.e., activity recorded in LMS) can be used as a predictor of students' academic performance (Moubayed et al., 2018), it does not necessarily mean that the students with the best performance will be the most active students (i.e., those that will read the learning materials most often). The comparison of the activity levels between clusters C1 and C3 shows that less-to-moderate successful students (C3) have been more active in LMS than the most successful students (C1), i.e. they were reading the materials more often. These results support the findings of Marques et al. (2018), who had a similar observation that the most successful students were not the ones with the highest average access scores within the e-learning platform. Likewise, we can conclude that students with the best final results are more confident in their domain knowledge, resulting in less frequent access to the course materials provided in LMS.

Cluster analysis results from Subsection 5.1.1 suggest that students have been using assessments more consistently than accessing the lessons. This is in line with the observations from Manwaring et al. (2017), stating that students' perception of the importance of an activity has a strong positive effect on students' engagement level, both cognitive and emotional. In line with this study, we can conclude that higher engagement in self-assessments gave our students better perception of learning and improvement.

Correlations observed in Subsection 5.1 would also suggest that students' activities (i.e., accessing the materials) did not contribute greatly to their understanding of knowledge domains, i.e. accessing materials does not significantly affect the increase of the knowledge because it does not imply that they have read and understood the materials. Although this is in contrast with findings from other studies (Orji & Vassileva, 2020; Nguyen, 2017; etc.), in our particular case it may be partially explained by two preliminary remarks:

a) The very technical nature of materials - materials can be printed-out from the LMS so students may have been reading them outside the LMS too.
b) The programmed learning paths of the course - realistically, students had to go through the lessons only once to unlock the access to self-assessments. So, once

they gained the access to self-assessments, which they arguably also perceive as more beneficial for their learning, they lost the interest to re-read the materials. Especially if they also had printouts for "easier" learning.

Decision to rely strongly on various assessment activities in the learning design used in this course is supported by Nguyen et al. (2017) and Lei et al. (2018), showing a significant relation between assessment activities and students' success rates, as well as the fact that both the learning design and the computer-based assessments affect students' online learning. Findings from Orji and Vassileva (2020) also suggest that the academic performance, as well as their final grade, can be predicted by students' engagement (i.e., activity recorded in LMS), assessment and assignment scores.

7 Limitations and future work

7.1 Study limitations

There were several constraints which have to be taken into account when trying to generalise the results of this study. As we have already mentioned, the study included a large number of students (200+) and therefore the results may not apply equally to situations where the ERT was conducted with smaller groups. It is especially seen in the communication activities between teachers and students, i.e. teachers cannot regularly communicate with every student to offer help or guidance. Another specific factor was the nature of self-assessment quizzes used in course. Since all the quizzes were automatically evaluated by LMS without teacher's intervention, all the questions were designed to enable the automatic grading (single/multi choice, matching, etc.) and the essay-type questions were not used. This limitation can also be linked to the large student groups - manual grading of essays would overburden the teachers, especially considering that all quizzes had an unlimited number of attempts. The nature of the course itself may also be a differentiating factor since the course used in this research is mainly theoretical and the results may not be generalised to more practical courses. Finally, since this study is the first stage of the planned research (see the possibilities for future research outlined above), only a limited set of criteria directly available in LMS and in the Faculty's Information System was measured - activity in reading lessons, results of self-assessment quizzes and the final grades. Other sets of criteria which could additionally explain students' behaviour were not included in this study (students' attitudes, learning styles and strategies, ICT literacy, studying conditions, etc.).

7.2 Future research

This study is focused on the so-called "specific domain of information" according to Anouar Tadlaoui et al. (2016) which involves the level of knowledge and other specific information about the learner such as records of learning activities and records of

evaluation. Since students enter this course in the first year of the undergraduate study programme, and since this is a rather complex and demanding 5 ECTS course, it usually results in a large number of failures in the end. Due to the fact that the high dropout rate can discourage students in further studying, there is a need to detect the students at risk at the early stages of their education and to assist them to master this course. Therefore, it was decided to focus on the data that show students' performance in the course since improving student retention rate is a vital objective for this particular group of students and it is tightly connected with the success rate of the study programme. However, according to the same source (Anouar Tadlaoui et al., 2016), user models can be enriched with the so-called "independent domain information" which involves users' goals, attitudes, motivation, background, experience and preferences. With this in mind, the next step in this research should include the latter data to better understand how different students' characteristics influence their learning and their final success. As the next step, a course satisfaction survey will be prepared with the socio-demographics questions that will enable further research on the higher failure rate of the DCs students.

Even without using enriched data sets, there are other possibilities for future research. For example, current cluster C3 (very active students, but struggling with studies and failing to achieve passing grade) would benefit from applying predictive analytics in order to enable early detection of such students and offering them more teacher guidance. Predictive analytics could also be used with current cluster C2 (inactive and failing students), again to enable their early detection and to incentivize their activity within the course.

8 Conclusion

This research follows contemporary innovations in education and uses different techniques of learning analytics (LA) to identify possible improvements in the course design and teaching strategies, as it is highlighted within the most recent research. It complements the field by determining actual student groups post hoc, based on the analysis of available data about students' activities throughout the entire semester and modeling students' knowledge. This is especially important in the current situation where emergency online teaching facilitated the shift to the online environment and a lot of data is recorded about students and their behavior that could be used to improve learning design and teaching strategies.

In this case, final grades were correlated with the results of two types of activities: (i) knowledge displayed on various online tests during the semester (many formative and summative online tests within the course - multiple flash exams and self-assessments, two midterms and the final exam) and (ii) frequency of accessing the course materials (i.e., students' activity).

Three different clusters of user groups were detected: C1 including students that earned passing grade, that were moderately active during semester and better understood the domain; C2 including the students that have failed the course, had low or no understanding of the domain and were mostly not active; and C3 including very

active students but many of them having difficulty with understanding the domain or having difficulties with their studies.

Finally, the research showed that the clustering of students can enable teachers to rethink about the design and teaching strategies used in the course to increase students' retention rate and their final grades. However, it was shown that in many cases the programmed learning path with well-balanced self-assessment can lead to successful mastery of the course and that many students can be guided automatically throughout the course using LMS thus helping teachers to better monitor students in large classes.

Abbreviations *ASW*: Average silhouette width; *BCP*: knowledge domain *Basic computer principles*; *CS*: knowledge domain *Computer software*; *CUC*: knowledge domain *Central unit of a computer*; *DC*: Dislocated study center; *ERT*: Emergency remote teaching; *HEI*: Higher education institution; *IOU*: knowledge domain *Input/Output unit*; *IS*: knowledge domain *Information systems*; *ISS*: knowledge domain *Information system security*; *LA*: Learning analytics; *LMS*: Learning management system; *MC*: Main center; *MU*: knowledge domain *Memory unit*

Authors' contributions Study conception, design and methodology received the most contribution from Igor Balaban and minor contribution from Miran Zlatović. Data collection, curation, analysis and visualization were performed by Danijel Filipović. Writing of the first draft of the manuscript, as well as all the intermediary versions and the final version was the collaborative effort of all authors. All authors read and approved the final manuscript.

Funding This work was co-financed by the Croatian Science Foundation project IP-2020-02-5071.

Data availability The datasets generated and/or analyzed during the current study are not publicly available due to GDPR restrictions i.e., them containing information that could compromise research participants' privacy but are available from the corresponding author on reasonable request.

Declarations

Competing interests The authors declare that they have no competing interests.

References

Anouar Tadlaoui, M., Souhaib, A., Khaldi, M., & Carvalho, R. (2016). Learner modeling in adaptive educational systems: a comparative study. *International Journal of Modern Education and Computer Science, 8*, 1–10. https://doi.org/10.5815/ijmecs.2016.03.01

Bignoux, S., & Sund, K. J. (2018). Tutoring executives online: what drives perceived quality? *Behaviour & Information Technology, 37*, 703–713. https://doi.org/10.1080/0144929X.2018.1474254

Bond, M., Bedenlier, S., Marín, V., & Händel, M. (2021). *Emergency remote teaching in higher education: mapping the first global online semester*. *18*. https://doi.org/10.1186/s41239-021-00282-x

Bontcheva, K., & Wilks, Y. (2005). Tailoring automatically generated hypertext. *User Modeling and User-Adapted Interaction, 15*, 135–168. https://doi.org/10.1007/s11257-004-5637-6

Britain, S. (2004). *A review of learning design:Concept, specifications and tools*.

Brusilovsky, P., & Millán, E. (2007). *User models for adaptive hypermedia and adaptive educational systems*. *4321*. https://doi.org/10.1007/978-3-540-72079-9_1

Camacho, D., Pulido, E., R-Moreno, M., Carro, R., Ortigosa, A., & Bravo, J. (2009). Automatic course redesign: Global vs. individual adaptation. *International Journal of Engineering Education*, *25*.

Chiang, M., & Mirkin, B. (2010). Intelligent choice of the number of clusters in K-Means clustering: an experimental study with different cluster spreads. *Journal of Classification, 27*, 3–40. https://doi.org/10.1007/s00357-010-9049-5

Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: a literature review for the last decade. *Expert Systems with Applications, 40*, 4715–4729. https://doi.org/10.1016/j.eswa.2013.02.007

Corrin, L., de Barba, P. G., & Bakharia, A. (2017). Using learning analytics to explore help-seeking learner profiles in MOOCs. *Proceedings of the Seventh International Learning Analytics &Amp Knowledge Conference*, 424–428. https://doi.org/10.1145/3027385.3027448

de Amorim, R. C., & Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences, 324*, 126–145. https://doi.org/10.1016/j.ins.2015.06.039

Dohn, N. B. (2010). Teaching with wikis and blogs: Potentials and pitfalls. *Proceedings of the 7th International conference on networked learning*, 142–150. https://www.lancaster.ac.uk/fss/organisations/netlc/past/nlc2010/abstracts/PDFs/Dohn.pdf. Accessed 2 June 2022.

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access: Practical Innovations, Open Solutions, 5*, 15991–16005. https://doi.org/10.1109/ACCESS.2017.2654247

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*, 304–317.

Francis, B., & Sasidhar Babu, D. (2019). Predicting academic performance of students using a hybrid data mining approach. *Journal of Medical Systems, 43*, 162. https://doi.org/10.1007/s10916-019-1295-4

Gilbert, B. (2015). *Online learning revealing the benefits and challenges*.

Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, M. (2020). The difference between emergency remote teaching and online learning. *Educational Review*. https://hdl.handle.net/10919/104648

Jia, J., & Zhang, J. (2019). *The analysis of online learning behavior of the students with poor academic performance in mathematics and individual help strategies* (pp. 205–215). https://doi.org/10.1007/978-3-030-21562-0_17

Kebritchi, M., Lipschuetz, A., & Santiague, L. (2017). Issues and challenges for teaching successful online courses in higher education: a literature review. *Journal of Educational Technology Systems, 46*, 4–29. https://doi.org/10.1177/0047239516661713

Korkofingas, C., & Macri, J. (2013). Does time spent online have an influence on student performance? Evidence for a large business studies class. *Journal of University Teaching and Learning Practice, 10*(2), 1–13.

Larsen, D. P., Butler, A. C., Aung, W. Y., Corboy, J. R., Friedman, D. I., & Sperling, M. R. (2015). The effects of test-enhanced learning on long-term retention in AAN annual meeting courses. *Neurology, 84*(7), 748–754. https://doi.org/10.1212/WNL.0000000000001264

Lei, H., Cui, Y., & Zhou, W. (2018). Relationships between student engagement and academic achievement: a meta-analysis. *Social Behavior and Personality: an International Journal, 46*, 517–528. https://doi.org/10.2224/sbp.7054

Lowenthal, P., Borup, J., West, R., & Archambault, L. (2020). Thinking beyond zoom: using asynchronous video to maintain connection and engagement during the COVID-19 pandemic. *Journal of Technology and Teacher Education, 28*, 383–391.

Manwaring, K. C., Larsen, R., Graham, C. R., Henrie, C. R., & Halverson, L. R. (2017). Investigating student engagement in blended learning settings using experience sampling and structural equation modeling. *The Internet and Higher Education, 35*, 21–33. https://doi.org/10.1016/j.iheduc.2017.06.002

Marques, B., Villate, J., & de Vaz, C. (2018). Student activity analytics in an e-learning platfom: Anticipating potential failing students. *Journal of Information Systems Engineering & Management*, *3*. https://doi.org/10.20897/jisem.201812

Martins, C., Faria, L., de Carvalho, V., & Carrapatoso, E. (2008). User modeling in adaptive Hypermedia Educational Systems. *Educational Technology & Society, 11*, 194–207.

Mazza, R., & Botturi, L. (2007). Monitoring an online course with the GISMO tool: A case study. *Journal of Interactive Learning Research*, *18*(2), 251–265.

Moubayed, A., Injadat, M., Shami, A., & Lutfiyya, H. (2018). Relationship between student engagement and performance in e-learning environment using association rules. *2018 IEEE World Engineering Education Conference (EDUNINE)*, 1–6. https://doi.org/10.1109/EDUNINE.2018.8451005

Muthuprasad, T., Aiswarya, S., Aditya, K. S., & Jha, G. K. (2021). Students' perception and preference for online education in India during COVID – 19 pandemic. *Social Sciences & Humanities Open, 3*, 100101. https://doi.org/10.1016/j.ssaho.2020.100101

Nguyen, L., & Do, P. (2008). Learner model in adaptive learning. *World Academy of Science Engineering and Technology, 45*, 395–400.

Nguyen, Q., Rienties, B., Toetenel, L., Ferguson, R., & Whitelock, D. (2017). Examining the designs of computer-based assessment and its impact on student engagement, satisfaction, and pass rates. *Computers in Human Behavior, 76*, 703–714.

Nguyen, V. A. (2017). The impact of online learning activities on student learning outcome in blended learning course. *Journal of Information & Knowledge Management, 16*, 1750040.

Orji, F., & Vassileva, J. (2020). Using machine learning to explore the relation between student engagement and student performance. *2020 24th International Conference Information Visualisation (IV)*, 480–485.

Otrel-Cass, K., Cowie, B., & Campbell, A. (2009). What determines perseverance in studying science? *Journal of Institutional Research, 14*(2), 30–44.

Rapanta, C., Botturi, L., Goodyear, P., Guàrdia, L., & Koole, M. (2021). Balancing technology, pedagogy and the new normal: Post-pandemic challenges for higher education. *Postdigital Science and Education, 3*, 715–742. https://doi.org/10.1007/s42438-021-00249-1

Rienties, B., & Toetenel, L. (2016). The impact of learning design on student behaviour, satisfaction and performance: a cross-institutional comparison across 151 modules. *Computers in Human Behavior, 60*, 333–341.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Ryabov, I. (2012). The effect of time online on grades in online sociology courses. *MERLOT Journal of Online Learning and Teaching, 8*, 13–23.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia, 126*, 1763–1768.

Schreck, J. (2003). *User modeling. In: security and privacy in user modeling* (Vol. 2). Springer. https://doi.org/10.1007/978-94-017-0377-2_2

Seel, N. (2012). Programmed learning. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (p. 2706). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_671

Smaili, E. M., Sraidi, S., Azzouzi, S., & Charaf, M. E. H. (2021). Towards sustainable e-learning systems using an adaptive learning approach. *Emerging Trends in ICT for Sustainable Development* (pp. 365–372). Springer.

Sun, A., & Chen, X. (2016). Online education and its effective practice: A research review. *Journal of Information Technology Education*, *15*.

Tseng, S. F., Tsao, Y. W., Yu, L. C., Chan, C. L., & Lai, K. (2016). Who will pass? Analyzing learner behaviors in MOOCs. *Research and Practice in Technology Enhanced Learning*, *11*. https://doi.org/10.1186/s41039-016-0033-5

Watson, S. L., Watson, W. R., Yu, J. H., Alamri, H., & Mueller, C. (2017). Learner profiles of attitudinal learning in a MOOC: an explanatory sequential mixed methods study. *Computers & Education, 114*, 274–285.

Wei, H. C., Peng, H., & Chou, C. (2015). Can more interactivity improve learning achievement in an online course? Effects of college students' perception and actual use of a course-management system on their learning achievement. *Computers & Education, 83*, 10–21.

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation, 81*, 2141–2155.