IJASSE | International Journal of Academic Studies in Science and Education

Year 2024, Vol. 2, No. 1, Pages 1-19 www.ijasse.com

OPEN ACCESS

Optimising Educational Outcomes: Data and Process Analysis Approaches with Attention to Self-Directed Learning

Bernadett Sarró-Oláh
Corvinus University of Budapest, Hungary

Szabina Fodor
Corvinus University of Budapest, Hungary

Article Info

Article History

Received:
17 April 2023

Accepted:
25 April 2024

Keywords

Learning analytics,
Educational process mining,
Higher education,
Self-directed learning

Abstract

The swift advancements in technology and the corresponding job market impose increasingly challenging and dynamic requirements on workers. This is a significant obstacle for higher education institutions in adequately preparing their students for contemporary expectations and equipping them to tackle future difficulties. Today's students are "digital natives", and they distinctly absorb knowledge and employ new strategies to learn compared to earlier generations. Hence, it is of utmost significance for higher education institutions to comprehend the student learning process. Learning management systems (LMS) can offer substantial assistance in this endeavor, as they facilitate comprehension of students' learning process, while log files also offer unbiased insights into individual adaptation. This study aims to investigate the learning mechanisms of Business Informatics students at Corvinus University of Budapest by analyzing Moodle's educational data. The objective of the study was to acquire a more comprehensive understanding of the learning patterns exhibited by students in higher education through the utilization of an extensive collection of log files. The central idea revolved around examining the behavioral, motivational, and interest-related dimensions of learning as indicators of self-directed learning. These were examined using two primary methodologies: data analysis and process analysis. The findings indicate that distinct learning patterns exist regarding data and learning processes. Additionally, there are variations in time management and information consumption habits. The results of this study have practical implications for identifying learning patterns and developing tailored interventions to enhance educational achievements.

To cite this article

Sarró-Oláh, B., & Fodor, S. (2024). Optimising educational outcomes: Data and process analysis approaches with attention to self-directed learning. *International Journal of Academic Studies in Science and Education (IJASSE), 2*(1), 1-19. https://doi.org/10.55549/ijasse.12

Corresponding Author: Bernadett Sarró-Oláh, bernadett.olah@uni-corvinus.hu

Introduction

The profound impacts of global megatrends such as Industry 4.0 and digital, COVID-19 and economic crisis, and conflict fundamentally alter our lifestyles, employment patterns, and social interactions. The increasing importance of adaptability and conscious response to continuous and complex change in individual, organisational, and societal development highlights the crucial role of learning. Learning enables young people and adults to engage actively and benefit from change rather than passively enduring it. By harnessing transformational capacities, learning empowers individuals to become agents of change.

As work skills requirements are undergoing significant changes, it is worth noting that one-fifth of adults lack the essential competencies necessary for daily life (World Economic Forum, 2023). This also highlights the deficiencies and issues with the current standardised education system. The calibre of education exerts a definitive influence on the calibre of society in its entirety, and a broad spectrum of research demonstrates a causal connection between the calibre of education and economic and social progress. Inadequate education has a negative impact on individuals, society, and the economy, but high-quality education can lead to substantial advancements. Investing in education yields benefits in policy domains beyond education, such as employment, health, and social inclusion.

Future Skills for the Workplace Demands and New Generation

Technological and economic transformations most visibly affect the employment market and working conditions. The advent of digitalisation and automation is not yet supplanting but profoundly altering how individuals engage in work. Conventional career trajectories will be interrupted. New employment and the evolving skill requirements of current jobs will necessitate frequent updating in low-status positions and high-skilled and high-prestige occupations. The labour market estimate for 2035 by the European Centre for the Development of Vocational Training (CEDEFOP) affirms that there will be a significant increase in demand for highly skilled people while the number of medium-skilled workers will remain relatively constant.

On the other hand, the occupation of low-skilled workers is expected to fall (European Centre for the Development of Vocational Training, 2023). According to the estimate provided by the Organization for Economic Cooperation and Development (OECD), millions of workers may soon need to acquire new skills and change occupations (OECD, 2023). The World Economic Forum's 2023 skills projection (World Economic Forum, 2023), which surveyed 803 organisations, predicts that by 2027, there will be a significant increase in the demand for creative thinking, analytical thinking, and technology literacy, in addition to industry-specific knowledge. Additional essential abilities encompass social and emotional attributes such as inquisitiveness and continuous learning; the ability to bounce back, adapt, and be nimble; and drive and self-awareness. This indicates that companies place significant importance on the adaptability of their employees, as well as their openness and eagerness to acquire new knowledge.

Education has a crucial role in enhancing human skills and competencies. Education can contribute to reducing inequality in the labour market by aligning learning results with the needs of the labour market and making training programs accessible to a broad audience. Digital technologies have the potential to either cause significant disruptions or bring about transformative changes in the labour sector. Multiple studies examine the elements that influence the transformation of job roles and offer a set of techniques to forecast future jobs. Meanwhile, digital technologies provide education establishments with options to utilize collaboration, e-learning, or other platforms to efficiently transfer knowledge. These platforms facilitate access to those who choose to study from any location, provided they possess suitable devices and internet connectivity. The COVID-19 pandemic expedited the digital transformation of the education sector and emphasised individual adaptation.

Self-Directed Learning

Nowadays, with the evolution of online platforms, the Learning Management Systems (LMS) used, such as Moodle, allow us to record a wide range of learner actions and interactions, from low-level events such as mouse movements and clicks to higher-level events such as learning paths based on self-directed learning (SDL) principles and variations in event frequencies.

Self-directed learning refers to the result of enabling learners to take charge of their own learning experience and make choices regarding the specific knowledge they choose to acquire and master (Boyer et al., 2014). SDL includes students strategically organising their approach to course tasks and activities, establishing objectives for completion, closely monitoring their progress and comprehension of the material, and critically evaluating their performance upon completing each task. (Hung et al., 2010). Moreover, self-directed learning is a significant dimension of online learning readiness (Sarro-Olah & Fodor, 2023), which means the students can apply SDL better the better they are at online learning, considering that all other things are constant. Li and colleagues have broken down SDL into four main components: self-planning, self-learning, self-evaluation, and self-reflection (Li et al., 2023). All events can be identified as part of an SDL dimension so that the student's digital presence can be used to assess the role of each category and determine where more effort should be made to develop online learning readiness.

Educational Process Mining

Educational Process Mining (EPM) enables the mapping of students' behaviour by tracking their navigation patterns and interactions with course content within a Learning Management System (Romero & Ventura, 2013). Therefore, educators can utilise the EPM to comprehend better students' learning patterns, the variables that impact their academic achievement, and the skills they acquire (Alqaheri & Panda, 2022). This enables them to construct and evaluate educational process models that accurately depict observed behaviour.

The models identified by EPM can be utilised to gain a deeper comprehension of the fundamental educational processes, to identify learning disabilities at an early stage, to generate personalised recommendations for students, to aid students with specific learning disabilities, and to provide feedback to students, teachers, or researchers, among other applications (Cerezo et al., 2020). Furthermore, the EPM enables the analysis of students' behaviours and the categorisation of these actions in relation to specific activities. Conformity analysis processes can be conducted to see whether a previously modelled behaviour aligns with the observed behaviour (Adams & Van Der Aalst, 2021). The EPM, as a methodology, has the potential to introduce novel approaches for studying students' learning behaviour and problem-solving skills (Cerezo et al., 2020; Romero et al., 2016).

The objective of this paper is to present the results of data and process mining procedures as possible tools for course development and individual support. The course observed was conducted in the Moodle platform, and the data was obtained from its event log. The paper is organised as follows: section II describes the methods used to conduct the experiments; section III presents the obtained results; section IV provides the discussions; and section V sets out the final considerations.

Method

The nature of the research is exploratory, so it was particularly important to define the main research directions, which were determined by the research objective. The main considerations were educational data mining and educational process mining, with regard to SDL and the practical usability of the results. In terms of methods, it was crucial to identify the appropriate data source, find the necessary tools and consider the methodology.

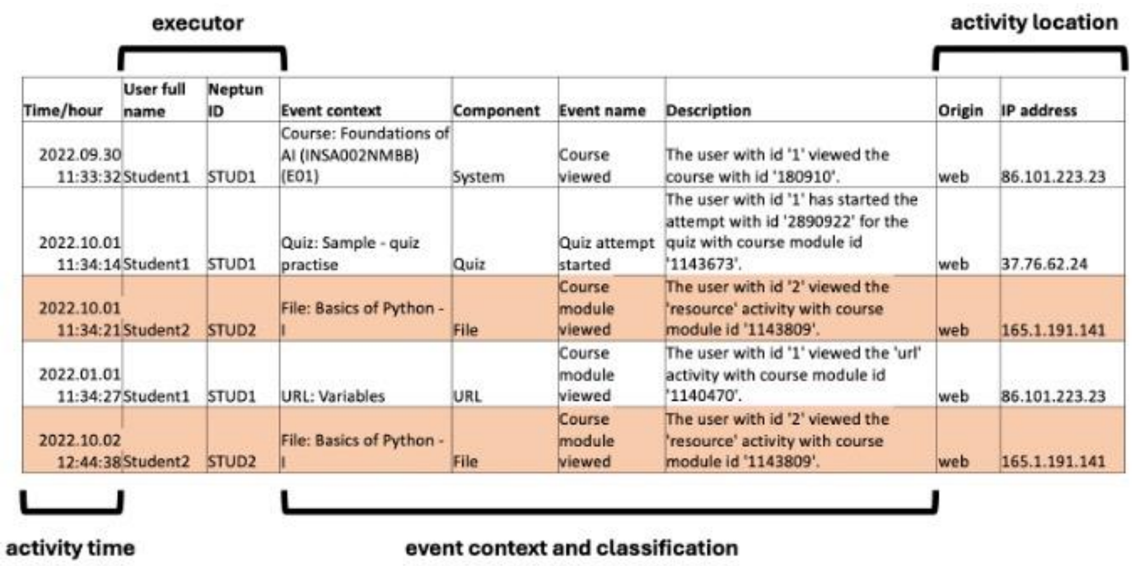
Participants

The study focuses on the learning log data of selected students at Corvinus University of Budapest, Hungary. These students were all relevant to the analysis since they participated in the Business Informatics programme Science, Technology, Engineering and Mathematics (STEM) courses and have experience using Moodle as an

LMS. Furthermore, these students were second-year undergraduates while the observation occurred in the fall semester of the 2022/2023 academic year; thus, due to the pandemic, they were used to face-to-face and online learning. The Fundamentals of Artificial Intelligence as the basis for the data collection course was chosen because of its practical nature. Students had to solve practical problems during the lessons, which they accessed via Moodle. As a result, 110 students' data were available. However, due to the individual curriculum and unusual circumstances, 98 students' data were reviewed in detail.

Data Collection and Pre-processing

The analysis of educational data is not a new field, but nowadays much more data can be collected, and activity tracking can provide an objective way of understanding students' learning and the use of digital tools provided by the university. Our primary source of data is Moodle, an LMS used in Corvinus University of Budapest; however, Neptun also offered information about belonging to the classes and some essential course or student information. Raw data includes details about the date and time of the activities, the identifier of the executors, the context and classifications of the events and the location characteristics of activities (see Figure 1)



Time/hour	User full name	Neptun ID	Event context	Component	Event name	Description	Origin	IP address
2022.09.30 11:33:32	Student1	STUD1	Course: Foundations of AI (INSA002NMBB) (E01)	System	Course viewed	The user with id '1' viewed the course with id '180910'.	web	86.101.223.23
2022.10.01 11:34:14	Student1	STUD1	Quiz: Sample - quiz practise	Quiz	Quiz attempt started	The user with id '1' has started the attempt with id '2890922' for the quiz with course module id '1143673'.	web	37.76.62.24
2022.10.01 11:34:21	Student2	STUD2	File: Basics of Python - I	File	Course module viewed	The user with id '2' viewed the 'resource' activity with course module id '1143809'.	web	165.1.191.141
2022.01.01 11:34:27	Student1	STUD1	URL: Variables	URL	Course module viewed	The user with id '1' viewed the 'url' activity with course module id '1140470'.	web	86.101.223.23
2022.10.02 12:44:38	Student2	STUD2	File: Basics of Python - I	File	Course module viewed	The user with id '2' viewed the 'resource' activity with course module id '1143809'.	web	165.1.191.141

Figure 1. Sample of original Moodle log file (Source: prepared by the authors)

The logging settings of Moodle are suitable for understanding an event by its existence and characteristics but not by its length since events include only the initial times of activities. The exact moments of completion, however, are not logged, hence, defining the precise duration of the activities is beyond the scope of this study. The first step was to prepare the log files for analysis, such as only events of the selected students were required. Other participants' activities or system logs were eliminated. Furthermore, only the activities during the course were of interest, so just the events between 29-08-2022 and 05-02-2023 were included. Research into possible duplication and system anomalies was also carried out at this stage. As a result, 54853 valid events were collected for the ninety-eight students surveyed. Python was used in the preparation process, given the extensive input data and multiple data sources. Another approach that made the use of this tool important was that big data raises key issues not only for processing but also for data privacy. In interpreting and using the data, the elements that would allow the identification of individual students were depersonalised, thus protecting the research participants. This process can be reliably tracked using Python.

Data and Process Analysis

The nature of the research justifies the use of a variety of tools and methods. Python was used for data analysis due to the reliability and transparency of the modules and procedures. Thus, after assisting with data cleaning, it has also proved useful for basic mathematical and statistical tasks, visualisation and hierarchical clustering. In addition, the RapidProm extension was applied for the process analysis procedure in RapidMiner. This tool offered an opportunity to examine events not only as a unit of activities but also as a link in a chain. Due to the specificity of log files, the Heuristic Miner algorithm was chosen. This algorithm can reveal the main elements of the process from the control-flow perspective and is a good choice because of its ability to deal with the noise in the data set and the possibility of identifying the main learning activity sequences (Weijters et al., 2006).

Results

Having clarified the source of the data and the tools needed, output data were examined according to two main criteria. These are gender differences and course activities. However, in the latter case, the time management aspect, SDL aspect, digital learning aspect and the learning process aspect are also discussed in more detail. The results show both the activity of the individuals studied and the useful information that can be obtained from the LMS, either during or at the end of the semester.

Gender Differences

Once the data has been pre-processed, detailed insight into student activity is possible. Data from ninety-eight students are examined, with gender differences being explored first. Activities were generated by seventy-five male and twenty-three female students, based on Moodle logs. Different amounts of activity were generated according to the different sizes of the gender groups, so an independent t-test can indicate similarities or differences between the two groups. Independent t-test emphasises that there is a statistically significant difference between male and female students in the average number of activities if the chosen significance level is 5% and both samples follow normal distribution (see Table 1).

Table 1. Results of the independent t-test by gender (Source: prepared by the authors)

	Male		Female	
	Mean	Stdev	Mean	Stdev
Number of activities	539.88	191.12	624.43	177.52

Interpreting the results, female students generated statistically more activity during the semester by the data observed.

Course Activities

This course supported obtaining practical knowledge about AI and parallel Python programming. Each week, the lecturer uploaded theoretical and supporting content and Python training codes, which students could use to try out the methods they had learned. The solutions to the practice codes were also uploaded a few days after the lessons. These contents provided the material for the practical exam at the end of the semester, so the way the students use it can be a guide for the instructor.

Time Management Aspect

An overview of the activities throughout the semester gives a comprehensive picture of the digital content used in learning. This provides insight into behaviour related to the timing of digital learning and the frequency of content use. Figure 2 highlights the evidence of a cyclical learning schedule in the present research.

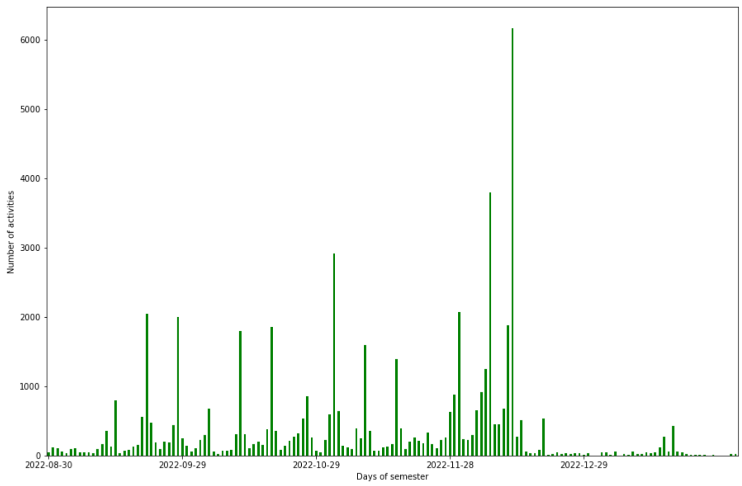


Figure 2. Number of student activities by the days of the semester (Source: prepared by the authors)

Students generated most of the events weekly on the day of the lessons. On the other hand, the exam was one of the most active, with most students taking the end-of-year test on 12-12-2022, for which they could and should have used Moodle.

In addition to the overall time usage, it is also possible to determine how many log files were generated by the students before assignments, how many before the exam and how many during other periods of the semester. The distance from these special days could categorise all events and could be counted per student. The event was considered part of an exam or assignments if it started 7 days before the exam. Finally, all ninety-eight students had standalone, close-to-exam and close_to_assignment events in individual composition. After data preparation, which included normality testing and standardisation, the data for each student became comparable. Based on the correlation values, it is possible to form clusters based on the three variables under study. Given the data and the lack of an ideal number of clusters, a hierarchical clustering procedure was chosen. Three clusters were the ideal choice based on the dendrogram (see Figure 3) and the agglomeration schedule.

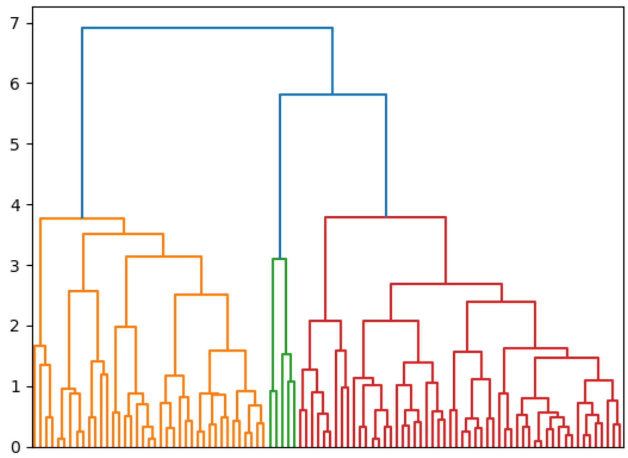


Figure 3. Dendrogram of hierarchical clustering (Source: prepared by the authors)

The first cluster, which includes fifty-four students, is called *Persistent digital learners*. According to the Moodle logs, these students showed higher than average activity levels during the semester, with no significant increase in activity levels before assignments and exams. The second group is the *Traditional learners*. Thirty-nine students belong to this group, and their main characteristic is that they were less active than average both during and at the end of the semester. The smallest group is *Deadline drive digital learners*, with five students. This cluster was active during the course but became particularly active around assignments, especially before the exam (see Figure 4). Therefore, the number and characteristics of the clusters suggest that a considerable proportion of students were active throughout the course, not just focused on the assignments and the exam.

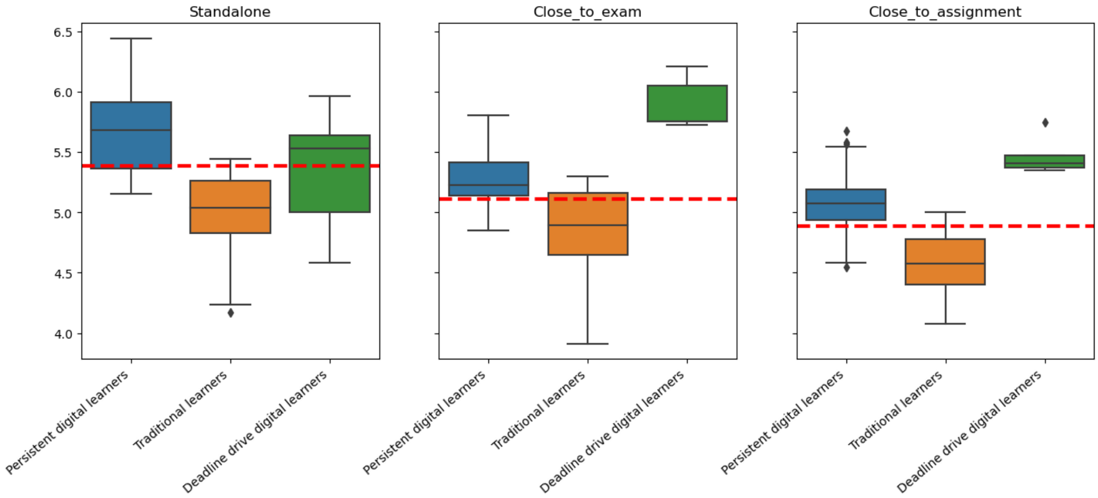


Figure 4. Boxplots about characteristics of clusters (Source: prepared by the authors)

Self-Directed Learning Aspect

Event logs from Moodle can be translated into SDL dimensions as well. Li and colleagues have presented their classification method in their work (Li et al., 2023), but the systematisation of events requires caution due to course specificities. In this study, therefore, the basic rules of classification are formulated according to the specificities of the own data. Events that focus on gathering information about courses and requirements are the parts of the Self-planning (SP) dimension. In contrast, the Self-learning (SL) dimension includes activities in which the reviewed person uses some tool in the learning process, such as presentations, practical supports, data sources and supplementary materials. In cases where the participant practises quizzes, prepares individually for the assignment or looks at the solutions for practical tasks or the lecturer's feedback, it can be classified as Self-evaluation (SE). The Self-reflection (SR) dimension includes events in which the students look at information related to the exam, or check the exam tasks, either repeatedly or retrospectively, and view the grades and related feedback.

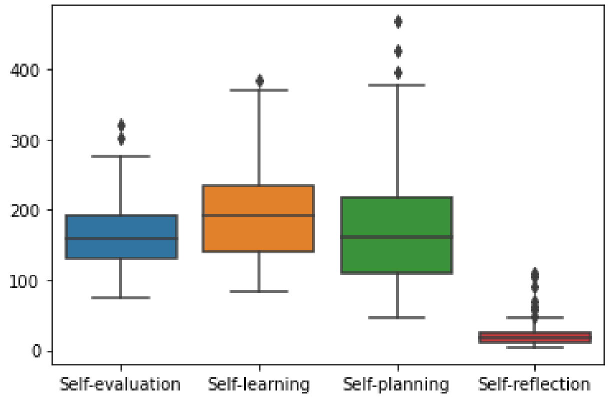


Figure 5. Boxplots about students' activities by SDL dimensions (Source: prepared by the authors)

Figure 5 highlights that the highest average number of activities among the SDL categories is found in the SL category and the lowest in the SR category. SP and SE are close to the average level of SL. The figure also shows that students found Moodle content useful mainly for self-study and were eager to gather information and evaluate their knowledge. This course, however, did not give enough space to the SR dimension or was not attractive enough for the students. From another aspect, students used the information available for the semester including by SP dimension. Furthermore, they also had access to learning support tools that are part of the SL dimension, which they used. Moreover, they had the opportunity to develop independently based on the solutions and individual feedback that came under the SE dimension. In terms of student behaviour, traces of SDL were recognisable, even though the use of Moodle was not compulsory, except for exam, quizzes and assignments.

Digital Learning Aspect

Categorising events can help to map activities from other aspects. Moodle has its classification for the events, namely components; however, this is not enough in some cases, and there are no built-in constraints to ensure collective understanding (Martínez-Carrascal et al., 2024). For a deeper insight, a classification is needed that points to the type of task in which the student's activity occurs (see **Hata! Başvuru kaynağı bulunamadı.**). In this case, the most important categories were some of the highlighted parts of the SL and SE dimensions, namely the presentation, the practice, the solution, and the supplementary types, as the instructor could also use them to assess the continuity of learning and possible individual obstacles even during the semester. The presentation-type activity was related to the number of times students viewed the uploaded theoretical background material. The Python code was available for exercise as a practical-type learning material, and the solutions could be viewed by clicking on the solution-type material. The supplementary-type resource provided various extra content on the topic, although, knowledge of it was not part of the requirement, but it provided an exciting insight for those interested.

Based on the Moodle log files, the reviewed students generated 3350 practice activities, 3204 presentation activities, 946 supplementary activities and 738 solution activities. There are different usage dynamics when looking at these events by relative month of the semester (see Figure 6Figure 6).

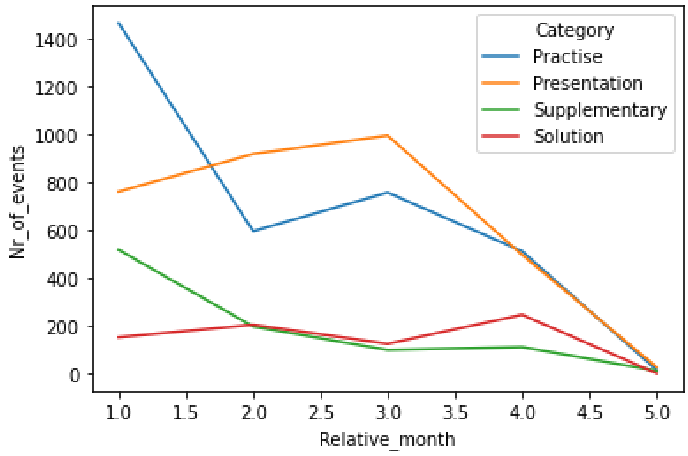


Figure 6. Number of events by the relative months of semester and the main categories (Source: prepared by the authors)

At the beginning of the semester, the most popular content was practice, with the number of recorded events decreasing as the course progressed. This is very similar to the supplementary type of event, which was, although, on a significantly smaller scale. Supplementary content was more popular at the beginning of the semester, but the amount of activity content decreased as the semester went on. The presentation-type events became more prevalent later in the semester since they reached their highest number of recordings in the third month. Solution content was also not popular in the first half of the semester, but as the exam approached, more activity was registered on this content.

Learning Process Aspect

The sequence identified by the student activities highlights the types of events that users prefer and how these actions are associated with each other. This approach reflects on the events generated during learning as a complex process. There are several possible tools, however, due to the specificity of the data, the Heuristic Miner algorithm was chosen. The heuristic process model provides a good solution for a detailed understanding of digital learning processes, as it is able to highlight important events and process relationships despite the noise in the input data (Weijters et al., 2006). The input data is taken from log files of Moodle and stored in a comma-separated value form. Only events belonging to the SL and SE dimensions were taken into account. The input data includes datetime attribute as a timestamp to the second, the masked Neptun IDs, which are the unique identifiers of the students, and the component classification categories, which are detailed in **Hata! Başvuru kaynağı bulunamadı.**. The RapidProm extension provides a tool to convert the input data into event log-type data. To do so, the event log factory was set to naive, the event classifier was defined as the event name, the trace identifier was specified as Neptun ID, and the event identifier was chosen as the category. Finally, after specifying the input data, the Heuristic Miner (Heuristic Net) was used to make the analysis executable. The fitness of the model is 0.88, calculated when the model is run. This value means that the model covers event logs well (Chanifah et al., 2021).

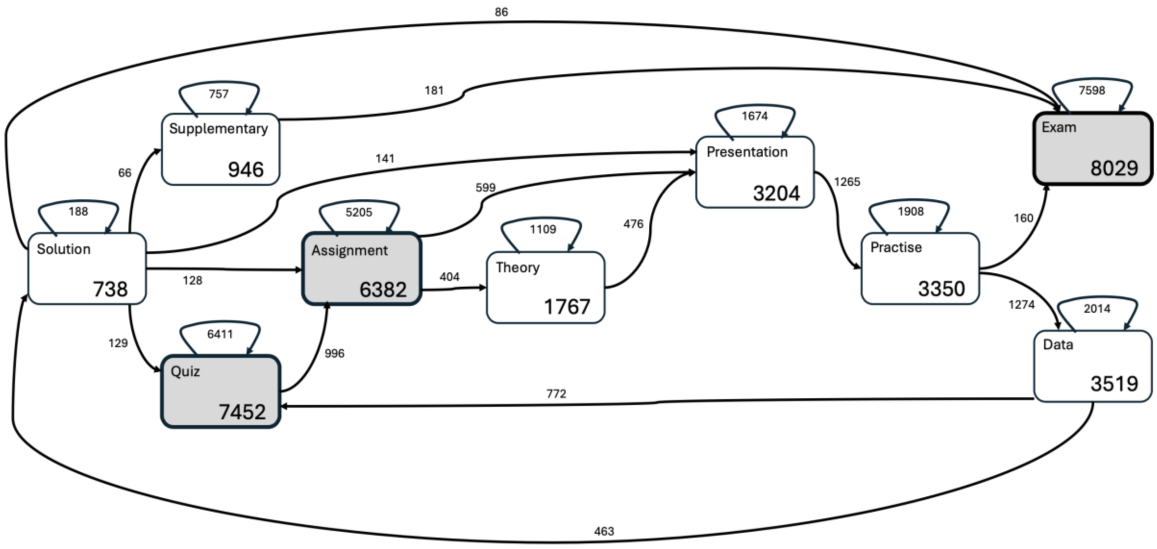


Figure 7. Heuristic process model of Moodle log events by cathegories (Source: prepared by the authors)

The results are presented in Figure 7. Categories, meaning event identifiers are marked with rectangles, showing the name and the frequency. Each category can return to itself in the process, indicated by a return arrow. The number on it shows the number of returns. This means that not only different events can follow each other, but also events of the same classification. In the case, however, where two different events are side-by-side, the direction of the edge between them indicates the direction of the relationship, and the frequency is also given next to the edges. The most frequent categories are those marked in grey. These were the compulsory ones to complete the course: the quiz, the assignment and the exam. It can be seen that 128 cases of solutions were followed by assignment events and 129 by quiz events, which suggests that students used solution content to prepare and achieve better results on the mid-term tests.

The presentation type content was followed by the practice type in 1,265 cases, so it can be said that the students had looked at the background knowledge needed to do the exercise before doing it. Practise type was followed 1274 times by data type activity, so the data used for the practical exercise was looked at by the students after the start of the exercise. And, whereas, data drives back to solution-type activity it highlights a sequence of practical learning. If the process is viewed from the exam side, the most common directions that precede it are practice, solution and supplementary. This can be explained by the fact that the final test was open book, so it can be seen that students used the previously uploaded content during the exam.

Discussion

The original log files were processed to extract information from the large amount of data that would support those involved in education. Considering all the activities of the course as a whole can be an approach to interpretation. In this case, the results can be used to target improvements to the course for the upcoming years. Nowadays, however, due to the wider educational representation of digital tools, meaningful data can also be extracted during the semester. Some of the methods could catalyse understanding by providing insights, such as identifying how participants schedule their learning, how they are ready to learn by themselves based on SDL dimensions or how they use the main content categories. In our study, practical content and presentations were the most common elements of digital learning, although the presence of these varied over the semester. In the process model, content related to the fulfilment of requirements, such as quizzes, assignments and exam, were the most frequently generated activities.

The results that Chanifah and colleagues presented in their work indicated similar results, as the most important activity types for the programming course they studied were quiz, course and URL (Chanifah et al., 2021). The course type mentioned in their case can be matched with the activity we identify as practice, and the URL can be data, practise, solution or supplementary type according to Table 2. In both cases, the quiz appears in the same way. The importance of computer-supported, real-time predictions has been mentioned before (Aldowah et al., 2019), however, in the context of an accelerated change of circumstances, its role has been enhanced. Our results may not be generalisable to all courses, but they provide supporting content that can be objectively extracted from Moodle log files.

An alternative approach is that students may also benefit from understanding the data, as the process model can be examined at the level of the individual. Our results have shown the learning process at the course level, but several studies have already stressed that it would be worthwhile to formulate personalised recommendations based on the EPM (Bogarín et al., 2018), and even individual learning pathways (Martínez-Carrascal et al., 2024).

In this study, we confirmed gender differences, which resulted in fewer female students participating in the course, but they were statistically significantly more active than their male peers, however, the present study did not find any reasons for this. The gender results are in line with the findings of Sáiz-Manzanares and colleagues on STEM courses, who have emphasized in their work that it is worth exploring ways of reducing the gender gap using learning data (Sáiz-Manzanares et al., 2021).

Classifying students by gender is not the only way to group them. In the present study, we distinguished three groups of students by their behaviour according to whether they used digital content near the exam, before assignments or during other periods of the semester. This is just one approach; other studies point out that grouping can also be done based on students' motivation and SDL level (van den Beemt et al., 2018). Furthermore, Li et al, focusing on online SDL, highlight that low and high-achieving groups in terms of performance show different learning patterns (Li et al., 2023). Additionally, students can be grouped not only by data but also by their individual learning processes (Ramos et al., 2021). This means that the data provide a wide range of opportunities to understand how students use digital tools in their learning, and the method chosen depends on the purpose. This research highlights only some approaches and does not look at offline learning. Moreover, we studied only a small group of students, and all participants were at Corvinus University in Budapest, Hungary. Furthermore, the behaviour of students taking a STEM course was observed. The characteristics of the course may also differ from other similar classes. Consequently, the log files of the students could be investigated more extensively in the future.

Conclusion

Concluding the results, the present research, based on the log files of Moodle as an LMS, sought to answer the question of how to use data and process analysis tools for understanding learning. First, the gender gap was analysed in the selected STEM course, showing that most of the participants, 75, were male and the minority, 23, were female. In contrast, it was proven through statistical independent t-test analysis that female students generated more activity on average than their male peers.

Looking at the whole semester, it can also be seen that students generated several log files weekly, which coincided with the contact hours. Between lessons, the activity rate is lower. Moreover, most of the activity is seen on the day of the exam when students were required to use Moodle to obtain grade. This evidence shows how students align their time management approach for requirements and, as seen in the case of the exam, it shows which events produce more activity. The students in the study adapted to the course and followed its pace in their individual learning.

In another approach, the use of time was also examined in terms of whether students studied mainly before exam, before assignments or during other parts of the semester. All students had all three of these activities, but in different combinations. According to these values, they were classified into three groups using a hierarchical clustering procedure. The first group is the Persistent digital learners with 54 students, who had higher than average activity throughout the semester, with no significant fallback. The second group is the Traditional learners. 39 students belong to this group and their main characteristic is that they were less active than average throughout the term. The smallest group is Deadline drive digital learners, with 5 students. In their case, the most important period was before assignments and exam, with average activity for the rest of the semester. Given the practical nature of the course under study, the expected behaviour would be to get as much activity from the students as possible. However, it is also necessary to mention that the extent of offline learning is not apparent from the data examined.

In terms of the readiness of students for self-directed learning, Moodle logs can also form a basis. In doing so, it was found that, on average, most of the activities that students completed were related to self-learning, even though its components were not mandatory for students to use. The self-planning dimension of the SDL was also significantly present, meaning that students used information that was shared on the online surface. Similarly, the self-evaluation dimension is also visible in the data, but this also includes some of the activities of the quizzes and the assignments, so a higher number of logs was expected. Self-reflection, as a collector of the exam and related content, showed the fewest events on average compared to the other activities. However, it can be said that non-mandatory content was also used by students, giving evidence of the presence of SDL.

Having collected the SL and SE content that were most relevant for individual learning in this course, we further investigated thes practice, solution, presentation and supplementary types of events. The two most numerous were the practice and presentation types. The former generated more interest in the first half of the semester and the latter in the middle of the semester. The solution and supplementary content types showed significantly less activity than the previous two types. The solution type tended to be more important in the month of the exam according to the logs, which can be explained by the preparation and by the fact that students could use previous solutions during the exam. Supplementary, compared to the practice-type content, showed more activity in the first half of the semester. It can be concluded that the most popular content was that which the students could use with the lecturer during the contact hours.

So far, results available with data mining tools have been presented, but these have not been able to identify activity types as sequences or at least directions of connections. Process mining therefore focuses on precisely this by including the SL and SE dimensions. Each activity could be followed by an event of the same classification, so describing the general process with the categories used, especially to fit all possible processes, would give an infinite number of outcomes. For this reason, and because of the specificity of the data, the Heuristic Miner algorithm was used to construct the process model. Process mining showed that the most frequently visited activities were related to the exam or any type of test. This result could be related to the fact that the visit of the named content was mandatory to complete the course. It can also be seen that the solution, practise and supplementary contents were in many cases followed by an exam-type event, meaning that students used the course content for preparation and even for the exam. Furthermore, a noticeable number of students followed the presentation category with the practise type, followed by data and finally the solution. This process shows that after reviewing the theoretical background, students used the Python codes to practice, then they opened the prepared data files and finally looked at the available solutions. Except for the exam and tests, this path is the most common direction, which was also the goal of the course.

Recommendations

In the present research, although we have studied a well-defined, small group, we make the following recommendations based on the results. First, it is worth considering at the course design stage how students' activities can and should be monitored. An important question is whether it is useful to interpret the data continuously throughout the semester, or if it is sufficient to do this only once at the end of the semester.
Based on the literature, it may be worthwhile to develop a recommended learning process, individual learning paths, or activity reports. In summary, the processing of log files can be useful for the instructor, who can intervene immediately if the data indicates that. For the students, it would provide direct feedback on their own activities and help them to follow their individual development paths. It is worth considering a structured form to which both students and teachers can have access. Processing log files can thus contribute to improvement.

Acknowledgement

Supported by the ÚNKP-23-3 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

References

Adams, J. N., & Van Der Aalst, W. M. P. (2021). Precision and Fitness in Object-Centric Process Mining. *2021 3rd International Conference on Process Mining (ICPM)*, 128–135. https://doi.org/10.1109/ICPM53251.2021.9576886

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, *37*, 13–49. https://doi.org/10.1016/j.tele.2019.01.007

Alqaheri, H., & Panda, M. (2022). An Education Process Mining Framework: Unveiling Meaningful Information for Understanding Students' Learning Behavior and Improving Teaching Quality. *Information (Switzerland)*, *13*(1). Scopus. https://doi.org/10.3390/info13010029

Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(1). Scopus. https://doi.org/10.1002/widm.1230

Boyer, S. L., Edmondson, D. R., Artis, A. B., & Fleming, D. (2014). Self-Directed Learning: A Tool for Lifelong Learning. *Journal of Marketing Education*, *36*(1), 20–32. https://doi.org/10.1177/0273475313494010

Cerezo, R., Bogarín, A., Esteban, M., & Romero, C. (2020). Process mining for self-regulated learning assessment in e-learning. *Journal of Computing in Higher Education*, *32*(1), 74–88. https://doi.org/10.1007/s12528-019-09225-y

Chanifah, S., Andreswari, R., & Fauzi, R. (2021). Analysis of Student Learning Pattern in Learning Management System (LMS) using Heuristic Mining a Process Mining Approach. *2021 3rd International Conference on Electronics Representation and Algorithm (ICERA)*, 121–125. https://doi.org/10.1109/ICERA53111.2021.9538654

European Centre for the Development of Vocational Training. (2023). *Skills in transitions: The way to 2035*. Publications Office. https://data.europa.eu/doi/10.2801/438491

Hung, M. L., Chou, C., Chen, C. H., & Own, Z. Y. (2010). Learner readiness for online learning: Scale development and student perceptions. *Computers and Education*, *55*(3), 1080–1090. https://doi.org/10.1016/j.compedu.2010.05.004

Li, Y., Jiang, Q., Xiong, W., & Zhao, W. (2023). Investigating behavior patterns of students during online self-directed learning through process mining. *Education and Information Technologies*, *28*(12), 15765–15787. https://doi.org/10.1007/s10639-023-11830-5

Martínez-Carrascal, J. A., Munoz-Gama, J., & Sancho-Vinuesa, T. (2024). Evaluation of Recommended Learning Paths Using Process Mining and Log Skeletons: Conceptualization and Insight into an Online Mathematics Course. *IEEE Transactions on Learning Technologies*, *17*, 555–568. https://doi.org/10.1109/TLT.2023.3298035

OECD. (2023). *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market*. OECD. https://doi.org/10.1787/08785bba-en

Ramos, D. B., Ramos, I. M. M., Gasparini, I., & De Oliveira, E. H. T. (2021). A new learning path model for e-learning systems. *International Journal of Distance Education Technologies*, *19*(2), 20–40. Scopus. https://doi.org/10.4018/IJDET.20210401.oa2

Romero, C., Cerezo, R., Bogarín, A., & Sánchez-Santillán, M. (2016). EDUCATIONAL PROCESS MINING: A TUTORIAL AND CASE STUDY USING MOODLE DATA SETS. In S. ElAtia, D. Ipperciel, & O. R. Zaïane (Eds.), *Data Mining and Learning Analytics* (1st ed., pp. 1–28). Wiley. https://doi.org/10.1002/9781118998205.ch1

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery*, *3*(1), 12–27. https://doi.org/10.1002/widm.1075

Sáiz-Manzanares, M., Marticorena-Sánchez, R., Muñoz-Rujas, N., Rodríguez-Arribas, S., Escolar-Llamazares, M.-C., Alonso-Santander, N., Martínez-Martín, M., & Mercado-Val, E. (2021). Teaching and Learning

Styles on Moodle: An Analysis of the Effectiveness of Using STEM and Non-STEM Qualifications from a Gender Perspective. *Sustainability*, *13*(3), 1166. https://doi.org/10.3390/su13031166

Sarro-Olah, B., & Fodor, S. (2023). Are Our Students Ready? Students' Online Learning Readiness in Higher Education Post-covid Era. In Y.-M. Huang & T. Rocha (Eds.), *Innovative Technologies and Learning* (Vol. 14099, pp. 326–335). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40113-8_32

van den Beemt, A., Buys, J., & van der Aalst, W. (2018). Analysing structured learning behaviour in Massive Open Online Courses (MOOCs): An approach based on process mining and clustering. *International Review of Research in Open and Distance Learning*, *19*(5), 38–60. Scopus. https://doi.org/10.19173/irrodl.v19i5.3748

Weijters, A. J. M. M., Aalst, van der, W. M. P., & Alves De Medeiros, A. K. (2006). *Process mining with the HeuristicsMiner algorithm*. Technische Universiteit Eindhoven.

World Economic Forum. (2023). *Future of Jobs Report 2023. Insight Report.* (p. 39).

Appendix

Table 2. Categories of Moodle activities (Source: prepared by the authors)

Category	Definition	Component (Moodle)
Assignment	Assignments requiring independent work, where students had to upload their solutions to the Moodle platform by a set deadline.	Assignment File File submissions Online text submissions Submission comments Assignment
Data	Prepared data files for practice and exercises.	File System URL
Exam	The final test of the semester.	Assignment File submissions Online text submissions Quiz System
Information	Information on deadlines, expectations and specific occasions to complete the course.	Assignment File Forum Quiz System

		User report
Practise	Practical exercises to solve each topic in Python.	File Quiz System URL
Presentation	The theoretical material that precedes the exercises, focuses mainly on the theory needed for the practical solution.	File System
Quiz	Tests at the beginning of the class, which check the students' current level of preparedness in 5 short multiple-choice questions.	Quiz System
Solution	Solutions to practical Python codes were published the weekend after the class.	System URL
Supplementary	Content that is not part of the requirement but helps to understand it.	File System URL
Theory	A more detailed theoretical overview than the presentation.	Folder System User tours

Author Information

Bernadett Sarró-Oláh
https://orcid.org/0009-0009-2940-3376
Corvinus University of Budapest
Fővám tér 13-15., 1093 Budapest
Hungary

Szabina Fodor
https://orcid.org/0000-0001-5459-7912
Corvinus University of Budapest
Fővám tér 13-15., 1093 Budapest
Hungary