Interdisciplinary Description of Complex Systems 23(5), 467-483, 2025

DATA AND PROCESS MINING IN ANALYSING STUDENT BEHAVIOUR

Snježana Križanić*, Katarina Tomičić-Pupek and Neven Vrček

University of Zagreb, Faculty of organization and informatics
Varaždin, Croatia

DOI: 10.7906/indecs.23.5.4
Regular article

Received: 30 April 2025.
Accepted: 1 September 2025.

ABSTRACT

The diversity of students' learning paths is crucial for acquiring knowledge. Although there are digital learning environments that provide many opportunities for managing the learning process, the rapid development of technologies can cause disruptions in the realisation of targeted engagement scenarios. Monitoring educational content use and increasing interaction frequency can contribute to better performance management and achievement of learning outcomes.

Data and process mining methods and tools play a significant role in the research of performance and deviations. Anonymized real data from one elective university course was collected and processed to create a dataset for the application of clustering and decision tree analysis in the KNIME Analytics Platform and for creating a process model in a process mining tool. The results show behavioural patterns for three clusters and provide insight into interaction types by identifying variables related to content engagement as effective discriminators for student grouping. The process model illustrates the diversity of engagement in choosing learning paths through the course (based on 51 cases performing 52 distinct activities with an average of 233 activities), while retaining the focus on the assignment deliverables. Insights obtained from the analyses are useful for the effective implementation of digital learning environments illustrating that no exceptional scenarios occurred in the course in terms of deviations in behaviour with the digital learning platform in relation to similar teaching and learning paradigms provided by the same teachers and that more interactive features combined with new technologies would be useful in providing more personalized learning paths.

KEY WORDS

data mining, clustering, decision tree, process mining, educational data

CLASSIFICATION

ACM: H33

JEL: D83, M00, O31

*Corresponding author, η: skrizanic@foi.unizg.hr; +385 42 390 893;
FOI, Pavlinska 2, HR – 42 000 Varaždin, Croatia

INTRODUCTION

Commitment to ensure engaged learning and provide feasible learning paths fitting various behavioural patterns is a desired feature while using digital learning environments. Teachers strive to develop a predictive model for improving teaching and course content management strategies based on real data. By using data on students' interaction and performance, teachers can discover learning patterns, supporting nearly personalized learning paths for students with different behavioural types. Data and process mining methods and techniques promise to contribute to revealing patterns and sequences, creating visualizations and analytics offering to predict student performance and deviations, expected use-cases, enabling data-driven education management. The expected aim of a more efficient education management is to provide insights for all stakeholders in how to design and maintain a satisfying customer journey (i.e., a student journey through a course) consisting of enough (but not too much) challenging (but not too hard to pass) activities, through delivering interaction touchpoints and resources enabling the acquisition of desired learning outcomes.

Recent developments in educational process mining have focused on discovering student learning paths and analyzing behavioral patterns within digital learning environments [1]. Educational data mining (EDM) and process mining have emerged as powerful analytical approaches for understanding and improving educational processes. The systematic application of data mining techniques in educational settings has been extensively documented, with researchers demonstrating the potential for analyzing student performance patterns and predicting academic outcomes [2]. The implementation of multimodal learning analytics has opened new avenues for understanding help-seeking behaviors and student interactions with both automated systems and human experts [3]. Comprehensive systematic reviews have confirmed that classification algorithms are most frequently applied in educational settings for evaluating student academic outcomes and identifying at-risk learners, while clustering techniques are commonly used for behavioural profiling and dropout prediction [4, 5]. These reviews emphasize that no single model uniquely predicts student performance, and the effectiveness of approaches depends heavily on data quality and contextual factors [5].

Although the motives for analysing student behaviour are often diverse, in this case the main goal was to investigate whether any exceptional scenarios occurred. This was done with consideration of the disruptive impact of new technologies on teaching and learning, and with a focus on what could be inferred from the measured engagement levels in one elective course at a higher education institution. The aim of this study is to explore student behavioural patterns in a digital learning environment using data and process mining techniques in order to identify engagement levels and detect potential deviations impacting course design and teaching strategies.

The article is organized in four main sections as follows. The Literature Review section identifies previous achievements in the field of data and process mining in education. The Methodology section describes the research design and procedures employed in this study. In the Data subsection, the dataset used for the research is presented in detail. Subsequently, the subsection Clustering and Decision Tree Procedure explains the application of data mining techniques, specifically clustering and decision tree analysis. Similarly, the Process Mining subsection describes the process mining approach applied in this study. The results are discussed in the Results section. The article concludes with the Discussion and Conclusion.

LITERATURE REVIEW

The reviewed literature demonstrates a growing interest in applying data-driven approaches, particularly clustering, classification, and process mining in order to analyse and improve

educational processes. Numerous studies employed process mining techniques such as process discovery and conformance checking to uncover students' learning paths, behavioural patterns, or system-level process inefficiencies. Clustering methods, especially k-means and Expectation-Maximization, were frequently used to group students based on their interactions or performance levels, while classification algorithms like Decision Trees, Naïve Bayes, and Support Vector Machines were applied to predict academic outcomes or identify at-risk learners. Unlike broader institutional studies, this work provides insights into how process mining can be tailored for smaller-scale, course-specific analysis to inform targeted pedagogical improvements.

Bey and Champagnat [1] explore how unsupervised learning methods, particularly clustering and process mining, can be applied to identify and analyse students' programming behaviours. The goal of the study was to discover behavioural profiles among first-year programming students and to examine the relationship between these behaviours and student performance in a university-level C++ course. The authors collected log data from 61 students using the Algo+ platform. Based on indicators such as the number of submissions, compilation errors, and average time between submissions, the authors applied k-means clustering to identify six behavioural profiles. These profiles were validated through ANOVA and ANCOVA tests, and the impact of behaviours on final exam scores was evaluated. Further, the authors used process mining techniques to analyse how students transitioned between behavioural clusters over time, differentiating between high-performing and at-risk students. The results showed that students who begin by carefully designing solutions (Cluster 1) tend to achieve better results, while those who rely heavily on trial-and-error strategies (Clusters 3 and 6) are generally associated with lower performance. The combination of clustering and process mining provided a comprehensive view of student learning paths and enabled the identification of critical behavioural patterns that influence success in programming courses.

Boztaş et al. [2] present a bibliometric analysis of the educational data mining (EDM) research field with a global perspective, aiming to analyse its performance, intellectual and social structure, and temporal development. The goal of the study was to investigate how EDM research has evolved over time, identify influential publications, authors, and countries, and discover emerging themes in the field. The results revealed a 1325% growth in publication output during the examined period, with "Computers & Education" as the most influential journal. The study highlights the growing interest in deep learning and emotional aspects of learning, suggesting these as promising future research directions. Additionally, the findings reveal limited international cooperation and a lack of theoretical grounding in many EDM studies, indicating areas for future improvement and development.

The study from Chen et al. [3] investigates differences in the help-seeking process of learners using either ChatGPT or a human expert by applying multimodal learning analytics and process mining. The research aims to understand how learners' help-seeking behaviours and sequences vary depending on the help source during an essay revision task. A lab experiment was conducted with 38 Chinese university students, divided into an AI Group (ChatGPT) and a HE Group (human expert), who completed an English essay writing and revision task. The authors applied both statistical analysis and process mining using the pMineR package to model and compare help-seeking behaviours between the groups. The results showed that the AI Group exhibited a non-linear help-seeking process, frequently skipping key metacognitive stages such as diagnosing questions and evaluating help, and often transitioned directly from asking for help to processing it. In contrast, the HE Group followed a more linear process aligned with established theoretical models.

Dol and Jawandhiya [4] present a systematic review of educational data mining (EDM) techniques used to analyse and predict student performance in educational settings such as

schools, universities, and e-learning platforms. The aim of the study was to examine EDM research trends, data mining algorithms, tools, and evaluation metrics. The study found that classification algorithms are the most frequently applied approach, often used to evaluate student academic outcomes and identify learners at risk. Other frequently used methods include clustering for behavioural profiling and dropout prediction, and association rule mining to discover learning patterns. Process mining was specifically used to analyse students' quiz-taking behaviour in learning management systems. The article concludes that although various approaches are used, no single model uniquely predicts student performance. Similarly, these authors provided a comprehensive systematic review and analysis of educational data mining (EDM) approaches used to predict students' academic performance [5]. The study aimed to identify and evaluate a wide range of data mining techniques, such as classification, clustering, and association rule mining, used in educational settings between 2010 and 2020. The results indicate that no single technique fits all educational contexts, and the performance of models depends heavily on the quality of input data and contextual factors.

The study published by dos Santos et al. [6] proposes a smart framework for performing risk and criticality analysis in industrial maintenance by integrating multicriteria decision-making (MCDM) methods and process mining techniques. The goal of the study was to develop a dynamic, flexible, and real-time evaluation model to support decision-making in maintenance planning by prioritizing industrial machines based on both qualitative and quantitative data. The results show that the proposed framework effectively identifies the most critical machines by integrating expert judgment and data-driven insights, with Machine 07 being consistently ranked as the highest priority for maintenance. The study highlights the feasibility of using a combined AHP–PROMETHEE approach with process mining to enhance maintenance decision quality in smart manufacturing environments.

Feng and Chen [7] present a comprehensive study that fully applies all three types of process mining: process discovery, conformance checking, and process enhancement. The aim of the research was to construct a complete educational process mining (EPM) framework using real event log data from a machine learning repository in order to analyse students' learning behaviours and provide actionable insights for students, educators, and administrators. Conformance checking revealed trade-offs between fitness and precision, with no single model achieving optimal results in both dimensions. Process enhancement was demonstrated using fuzzy instances and animations to visualize high-performing student behaviour, suggesting that successful learners tend to repeat and evenly distribute their activities. The study concludes that process mining can transform raw educational log data into actionable knowledge that supports personalized learning paths, improved teaching strategies, and data-driven education management.

The paper from Ghazal et al. [8] presents a systematic literature review that focuses on identifying and analysing empirical case studies where process mining techniques have been applied within the educational domain. The objective of the study is to assess the current state of Educational Process Mining (EPM), highlight its potential benefits, and provide future research directions. Results indicate that the majority of case studies (97%) employed the discovery type of process mining, with ProM being the most commonly used tool (84%). Clustering techniques were frequently used to handle unstructured and heterogeneous logs. Methodologies such as the Process Diagnostics Method (PDM) and various ad-hoc approaches were identified, although no standardized, domain-specific EPM methodology currently exists. The review emphasizes key challenges in EPM, including handling voluminous and heterogeneous data, interpretation of results, and a lack of repeatable implementation frameworks.

Grigorova et al. [9] provide a conceptual overview of data mining and process mining techniques and their application in educational environments, with a focus on improving traditional and e-learning processes. The aim of the paper is to describe the principles, methods, and tools of data mining and process mining, and to highlight their potential for enhancing educational systems through better data utilization. The paper discusses various data sources used in educational data mining, including learning management systems, intelligent tutoring systems, and collaborative learning platforms, and explores the types of problems these methods address like predicting student performance.

The article from Hicheur et al. [10] presents a distributed computation platform designed for educational process discovery and analysis. The goal of the research is to provide a scalable, flexible, and interactive system that allows educational institutions to analyse large-scale event logs using advanced process mining and data mining services. The authors propose an architecture based on an Enterprise Service Bus that integrates data sources, analysis tools, web portals, and service connectors, allowing real-time feedback and performance enhancement in educational environments. The platform supports process model discovery, conformance checking, and process enhancement, and enables run-time personalization of educational processes through features such as course recommendations and violation detection. A comparative study was conducted on several clustering techniques used to partition large educational event logs for process mining purposes, including sequence clustering, trace clustering, and Disjunctive Workflow Schema. Sequence clustering was identified as the most effective approach due to its ability to produce readable and simplified process models from large logs. The study concludes that distributing process mining tasks across multiple computing nodes improves performance and scalability.

The paper from Intayoad et al. [11] proposes a method to enhance personalized learning by discovering and analysing students' learning paths using process mining techniques. The objective of the study is to extract insight into students' actual learning processes based on their different learning styles and characteristics. The discovered process models were evaluated using the fitness replay metric to assess how well the models represent actual behaviour. Although the models achieved relatively low fitness values due to the noisy and less-structured nature of human learning behaviour, they were able to visually present full learning processes and identify variants among different student groups.

Juhaňák et al. [12] explore student behaviour patterns in online quiz-taking activities within learning management systems, with a particular focus on process mining as a method to uncover behavioural sequences. The aim of the study is to analyse students' interaction patterns during quiz-based activities in Moodle and to determine whether specific behavioural types can be identified and classified. The research applies a process-oriented approach by using process mining techniques to model and evaluate students' quiz-taking behaviour across multiple courses and quiz settings. The study reveals that students follow distinct strategies, and the process models offer a visual and analytical way to distinguish among these behavioural types. The results suggest that process mining can serve as a valuable tool not only for educational research but also for improving the design of online assessments and identifying students who may be at risk or using ineffective learning strategies.

Levin [13] presents a combined approach of process mining and expert feature engineering to predict students' efficient use of time during high-stakes computer-based assessments. The goal of the study was to develop a predictive model using students' interaction logs from the NAEP 8th grade mathematics test and determine whether students would manage their time effectively in Block B of the exam based on their behaviour in Block A. The methodology involved generating a large number of features (approximately 330) from log data using both expert-engineered metrics: focused on frequency and time-related patterns, and process mining

techniques that identified common action sequences. Feature importance analysis showed that time spent on specific items and average time on repeated sequences were key predictors. The results confirmed that combining process mining with expert feature engineering could yield interpretable features that significantly contribute to modelling time management behaviour in assessments.

The article from Ramaswami et al. [14] investigates the effectiveness of educational data mining techniques in predicting students' academic performance and explores whether the integration of process mining features can enhance the accuracy of such predictions. The study is based on real-time and self-paced interaction data collected via Xorro-Q, a web-based classroom engagement tool, in an engineering course. The dataset included 240 students and comprised in-class and out-of-class participation data, assessment scores, and prerequisite course results. Process mining was used to extract behavioural features through process discovery (using the Inductive Miner) and conformance checking in the ProM framework. The authors conclude that a long-term institutional commitment to learning analytics and access to larger historical datasets would strengthen the predictive capacity of EDM models and support early identification of at-risk students.

Romero et al. [15] introduce a practical tutorial and case study combining clustering and educational process mining using Moodle log data to improve model clarity and insight into student learning behaviour. The goal of the study is to propose a method that enhances the interpretability and performance of EPM models by grouping students with similar characteristics before applying process mining. The dataset includes Moodle interaction logs to discover behavioural process models. The discovered models for different student groups showed varying levels of complexity and fitness. The analysis revealed that passing students exhibited richer and more diverse interaction patterns with quizzes and forums, while failing students showed limited engagement. The authors conclude that clustering prior to EPM improves model comprehensibility and accuracy, and that these models can help instructors identify at-risk students and tailor interventions.

The article from Salazar-Fernandez et al. [16] investigates the impact of losing a need- and merit-based scholarship on students' curricular progress, dropout rates, and graduation outcomes, using a curricular analytics approach based on process mining. The results show that students who lost the BS experienced significantly slower curricular progress and higher dropout rates compared to those who maintained it. Further on, students who switched to government loans after losing the BS had better outcomes than those who became self-funded. CART models revealed that cumulative academic progress at the point of scholarship loss, gender, and academic field were key predictors of final outcomes.

Simić et al. [17] present a case study on enhancing project-based learning in higher education through data-driven analysis and visualisation using dashboards. The objective of the research is to support lecturers in supervising digital project-based courses by leveraging behavioural data captured from digital tools such as Jira, Confluence, and Mattermost. Evaluation of student behaviour using indicators such as Jira ticket lifecycle transitions revealed that the dashboards enabled timely feedback and behavioural adjustments, for example, improving the proper usage of ticket statuses. The results demonstrate that integrating data-driven dashboards into project-based courses positively impacts feedback quality, student engagement, and learning outcomes.

The article from Wibawa et al. [18] presents a systematic literature-based study on the use of learning analytics and educational data mining to enhance science and technology learning. The goal of the study is to define the concepts of learning analytics and data mining, outline their commonly used methods, and discuss their applications in educational contexts,

particularly within learning management systems. The study identifies a broad range of data mining methods relevant to learning analytics, including classification and prediction, clustering, outlier detection, relationship mining, social network analysis, process mining, and text mining.

The review from Yunita et al. [19] presents a systematic literature review on the use of big data for learning analytics and educational data mining, with the aim of identifying current research trends and proposing future directions for developing intelligent automation systems in education. The authors followed Kitchenham's [20] methodology for systematic literature reviews, filtering an initial pool of 480 papers down to 42 empirical studies published between 2016 and 2020. The review found that most studies used learning log data and focused on improving the learning process, profiling students, improving student retention, and evaluating student feedback. Techniques such as classification, clustering, regression, association rules, and text mining were frequently applied. Clustering was commonly used to identify student behavioural patterns, and classification methods such as Decision Trees, Naïve Bayes, and Artificial Neural Networks were used to predict academic performance or dropout risk.

A common finding is that combining multiple techniques (like clustering followed by process mining) often leads to more interpretable and actionable insights. Several studies also highlighted the importance of using real-time or multimodal data from platforms such as Moodle [12, 15], Jira [17], or Xorro-Q [14], with some research incorporating multimodal approaches that combine multiple data sources [3]. Despite methodological differences, the literature collectively underscores the value of process-oriented analysis in understanding learning behaviour and enhancing educational decision-making.

The reviewed literature reveals several key findings that motivate the current research. First, while comprehensive frameworks for educational process mining exist [7,8], most studies focus on large-scale implementations or standardized learning environments, with limited attention to smaller, course-specific contexts where exceptional behavioral scenarios may be more easily identified. Second, although the combination of clustering and process mining has proven effective for understanding student learning paths [1, 15], there is a gap in research specifically targeting the detection of unusual engagement patterns that deviate from typical behavioral trajectories. Third, while systematic reviews confirm the effectiveness of various EDM techniques [4, 5], they also emphasize that no single approach fits all educational contexts, suggesting a need for context-specific methodological adaptations. Finally, despite advances in real-time and multimodal data analysis [3, 12, 14, 17], there remains limited research on how process mining can be applied to identify exceptional scenarios in higher education elective courses, particularly in the context of technology-enhanced learning environments. These gaps in the literature provide the foundation for study's focus on exploring student behavioral patterns to detect potential deviations that could inform targeted improvements in course design and teaching strategies.

METHODOLOGY

Having in mind the variety of approaches from the literature review, this research builds upon the need of combining several techniques to gain more insights into learning behaviour, adding on the methodological framework proposed by Križanić [21], in which cluster analysis and decision tree techniques were applied to educational log data in order to explore student behaviour in an e-learning environment. In that study, a structured approach to educational data mining was proposed and illustrated through a step-by-step process including following phases: dataset selection, log file extraction from the e-learning system, data cleaning to remove irrelevant information, data partitioning for relevant feature extraction, clustering to identify behavioural student groups, and finally, classification through decision trees to further interpret

the identified groups. This study adopts the same general procedure, adapting it to a new dataset collected from a different academic generation and a different course. The analysis is conducted using the KNIME Analytics Platform instead of RapidMiner. KNIME was chosen due to its user-friendly, visual programming interface that allows intuitive design and execution of complex data workflows without the need for extensive coding. Its wide range of built-in nodes for data pre-processing, machine learning and visualization makes it particularly suitable for educational data mining tasks. By following the structured methodology, the aim is to ensure methodological consistency with previously validated approaches. Adding on the methodological framework from [21], process mining was performed on the same set of data used for data mining techniques to gain more insights into learning behaviour of students.

DATA

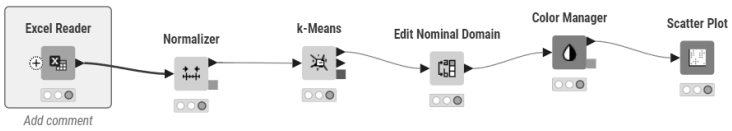
The dataset used in this study was extracted from anonymized log data collected from the Moodle-based e-learning system of a course focused on the development of process-oriented applications. It included records from 51 students enrolled in the course, covering the period from the beginning of October 2024 to the beginning of February 2025. During this time, students collectively performed an average of 233 interactions with the e-learning system, with the number of activities per student ranging from 61 to 896. The original data consisted of detailed timestamped interaction records reflecting students' activity in the digital learning platform throughout the academic semester. In order to ensure compliance with ethical standards and data privacy regulations, all personally identifiable information was removed. Each student was assigned a unique anonymized identifier (e.g., Student_001, Student_002), and records associated with course teachers or system administrators were excluded from the analysis. The collected log data included multiple dimensions of student activities such as course access, file access, forum engagement, user report views, and module activity views. To facilitate further analysis, these raw logs were cleaned and structured. Activities were categorized and grouped based on their context (e.g., file accesses, forum accesses, system usage), enabling the extraction of meaningful behavioural indicators for each student. This grouping was performed to reduce noise in the dataset and to generate aggregated features suitable for educational data mining.

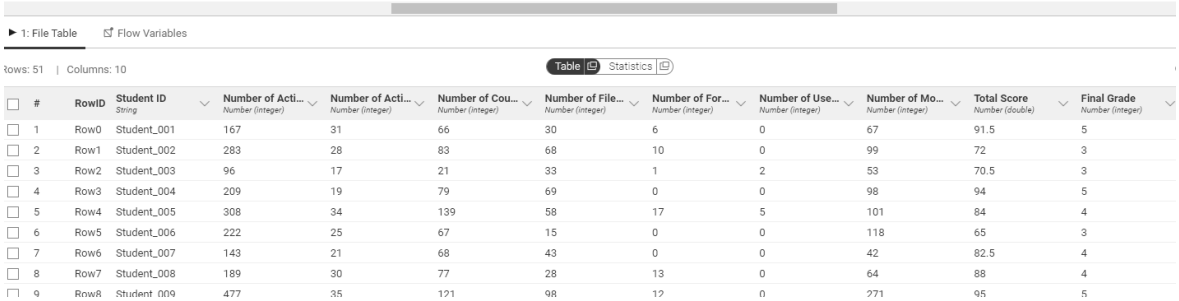
In addition to behavioural log data, student performance data were integrated from a separate source, containing each student's total score and final grade for the course. The datasets were merged based on the anonymized student IDs, resulting in a unified dataset that combines engagement metrics with student outcomes. The final version of the dataset, used for analysis, contains (per each student) variables such as the number of activities, number of active days, course access frequency, files and forum usage, module views, and the corresponding total scores and final grades. This processed dataset served as the foundation for the application of clustering and decision tree analysis in the KNIME Analytics Platform.

CLUSTERING AND DECISION TREE PROCEDURE

The clustering analysis was performed using the k-means algorithm in KNIME, based on a dataset containing aggregated indicators of student activity (Figure 1). The procedure started with the Excel Reader node, which imported the anonymized dataset in .xlsx format containing 51 student records and 10 features. This included activity-related attributes such as number of file accesses, course views, forum interactions, and the final score and grade achieved by students. Prior to clustering, the selected input attributes (Number of Activities, Number of Active Days, Number of Course Accesses, Number of File Accesses, Number of Forum Accesses, Number of User Report Views, and Number of Module Views) were normalized using the Min-Max normalization algorithm (applied via the KNIME Normalizer node). This algorithm scales each feature to a [0, 1] range, ensuring comparability of variables and

eliminating bias due to differing value ranges. Target attributes such as Total Score and Final Grade were explicitly excluded from normalization and clustering to avoid bias.





▶ 1: File Table | Flow Variables

Rows: 51 | Columns: 10 — Table | Statistics

	#	RowID	Student ID (String)	Number of Acti... (Number (integer))	Number of Acti... (Number (integer))	Number of Cou... (Number (integer))	Number of File... (Number (integer))	Number of For... (Number (integer))	Number of Use... (Number (integer))	Number of Mo... (Number (integer))	Total Score (Number (double))	Final Grade (Number (integer))
☐	1	Row0	Student_001	167	31	66	30	6	0	67	91.5	5
☐	2	Row1	Student_002	283	28	83	68	10	0	99	72	3
☐	3	Row2	Student_003	96	17	21	33	1	2	53	70.5	3
☐	4	Row3	Student_004	209	19	79	69	0	0	98	94	5
☐	5	Row4	Student_005	308	34	139	58	17	5	101	84	4
☐	6	Row5	Student_006	222	25	67	15	0	0	118	65	3
☐	7	Row6	Student_007	143	21	68	43	0	0	42	82.5	4
☐	8	Row7	Student_008	189	30	77	28	13	0	64	88	4
☐	9	Row8	Student_009	477	35	121	98	12	0	271	95	5

Figure 1. Clustering and file table properties of the Excel Reader node.

The normalized data was then passed to the k-Means node, where the number of clusters was set to three ($k = 3$). The clustering was performed with random initialization, using a static seed for reproducibility, and the maximum number of iterations was limited to 99. Following the clustering, the Edit Nominal Domain node was used to transform the resulting cluster labels (cluster_0, cluster_1, cluster_2) into nominal values. This allowed the Color Manager node to assign a colour to each cluster for visualization purposes. Finally, the Scatter Plot node was configured to display, for example, the Number of File Accesses on the x-axis, Final Grade on the y-axis, and Cluster as the colour dimension. The result was a coloured visualization of student distribution across clusters, which illustrated clear differences in behavioural patterns related to learning resource usage.

In the second phase, a classification model was developed using a Decision Tree Learner to further analyse the characteristics of student clusters. The following Figure 2 shows a decision tree model created in KNIME.

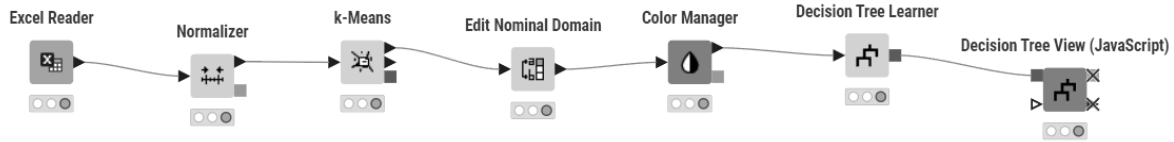


Figure 2. Decision tree model.

The model used the previously assigned Cluster attribute as the class column, which became the target variable in this step. The same pre-processed dataset was used as input, with normalized student activity indicators serving as predictor variables. Within the Decision Tree Learner, the Gini index was selected as the quality measure, and no pruning was applied. Reduced error pruning was enabled, and the minimum number of records per node was set to two. The algorithm was configured to automatically determine the average split point and used 16 threads for parallel processing. The output of the decision tree model was visualized through the Decision Tree View (JavaScript) node. This allowed for an interactive view of the tree structure, where the hierarchical splits provided insight into which activity features were most

influential in predicting cluster membership. The root and internal nodes of the tree were automatically determined based on optimal splits, and the final structure offered interpretable rules describing typical behavioural patterns within each cluster.

PROCESS MINING

In addition to data mining, process mining was performed on the same set of data for getting a comprehensive view of student learning paths. Anonymized data about activities (i.e., features of the Moodle systems) performed by students across the digital learning platform. Based on the data describing students' interactions with the learning platform throughout the semester, the data set was imported into Celonis, a process mining tool [22] suitable for generating process models and case analysis. Based on the data consisting of student ID (used as a case ID for generating a process model), activity name (i.e., the feature used in the learning platform) and timestamps within almost 12 000 entries a process model describing students' behaviour was created.

RESULTS

After analysing the clustering results of 51 students based on their activities within the digital learning system, three distinct clusters emerged: cluster_0, cluster_1, and cluster_2, each reflecting different patterns of engagement and behaviour. Cluster_1 is the largest group, comprising students who demonstrate moderate engagement across various learning activities. These students maintain consistent interaction with the platform's resources, resulting in generally good academic performance with most students achieving grades 3-5 (on a scale from the lowest 1 to the highest 5), though some receive 3s. Cluster_0 represents highly engaged learners characterized by notably higher activity values, particularly in forums and content interactions. Students in this cluster typically achieve excellent academic outcomes, predominantly receiving final grades of 4-5, with many earning the maximum grade of 5. Their behaviour suggests proactive, self-motivated learning with comprehensive utilization of available resources. In contrast, Cluster_2 contains students with generally lower engagement metrics, particularly showing minimal interaction with certain platform features. While there are exceptions (like Student_010 with a high grade despite being in Cluster_2), this group tends toward lower total scores and more variable final grades. The clustering effectively differentiates student engagement patterns, with activity levels in forums and content interaction serving as particularly distinctive features between the clusters, ultimately correlating with student performance outcomes.

The scatter plot from Figure 3 visualizes the relationship between students' final grades and their number of file accesses, with data points color-coded according to cluster membership. The x-axis represents the normalized number of file accesses, while the y-axis shows the final grade, ranging from 1 to 5. Students belonging to cluster_1 (red) are distributed mostly in the mid-range of file access frequency and are associated with a wide range of final grades, predominantly from 3 to 5. Members of cluster_0 (green) tend to have higher numbers of file accesses and consistently higher final grades, indicating a possible link between high engagement with learning content and academic success. Students in cluster_2 (brown) exhibit moderate file access levels but are less frequent and generally associated with lower to moderate grades.

The scatter plot from Figure 4 illustrates the relationship between the number of activities completed by students and their final grades, with data points differentiated by cluster membership. The x-axis represents the normalized number of completed activities, while the y-axis indicates the final grade. Students from cluster_1 are predominantly concentrated in the lower to middle range of activity completion, yet they achieve a wide spread of final grades,

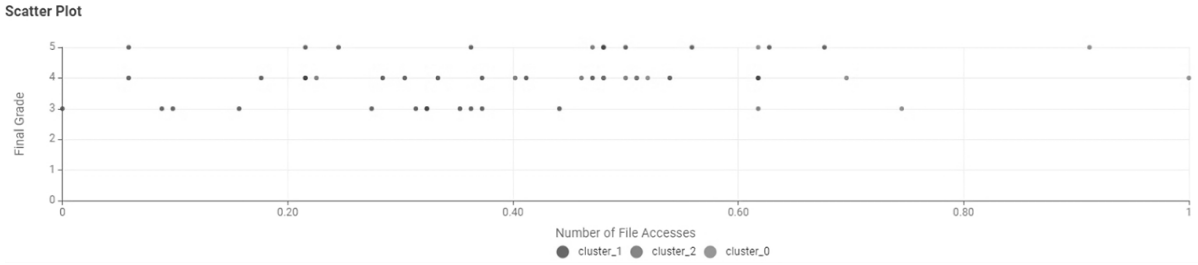


Figure 3. Clusters showing the relationship between final grade and number of file accesses.

mostly between 3 and 5. Students in cluster_0 generally exhibit higher numbers of completed activities and are associated with the highest final grades, reinforcing the pattern of strong engagement correlating with academic success. Cluster_2 students are fewer in number, situated in the mid-range of activity completion, and tend to achieve average final grades. The plot suggests that while moderate activity levels can result in good academic outcomes, students who consistently perform a higher number of activities (particularly those in cluster_0) are more likely to achieve top grades.

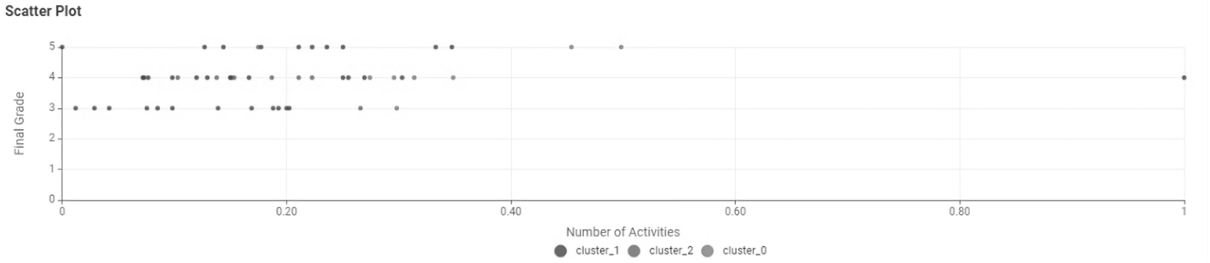


Figure 4. Position of clusters based on final grade and number of activities.

Analysis of the scatter plots further reveals the distribution of students across clusters. In both visualizations, the majority of students (over 30 individuals) belong to cluster_1, which represents those with moderate engagement. In contrast, the smallest group, with only 7 students, is cluster_2. This distribution underscores the predominance of moderately engaged students within the student sample, as well as the relative scarcity of students exhibiting either very low or very high engagement patterns.

The decision tree visualized in the Figure 5 represents a classification model generated in KNIME based on the previously obtained clustering results. The model was trained using the "Decision Tree Learner" node with the cluster variable as the target class column and the Gini index as the chosen quality measure.

At the root of the tree, the first and most informative attribute for cluster differentiation is the number of forum accesses. Students with a normalized value less than or equal to 0.39 are mostly assigned to cluster_1 (92,1%), indicating moderate or low forum engagement is characteristic of this group. This node further splits on the number of file accesses, where students with values above 0,69 are classified into cluster_0 (100%), indicating that high file access is a strong predictor for the most engaged cluster. The left branch of this split confirms that the vast majority of remaining students (97,2%) remain in cluster_1. On the other side of the tree, students with forum access values greater than 0,39 are further split based on their number of course accesses. If this number is less than or equal to 0,41, the majority of instances (88,9%) fall into cluster_2, which is associated with low to moderate overall platform interaction. When course access goes beyond this limit, all instances are classified into cluster_0. This confirms that students who are highly active in different ways are more likely to belong to this group.

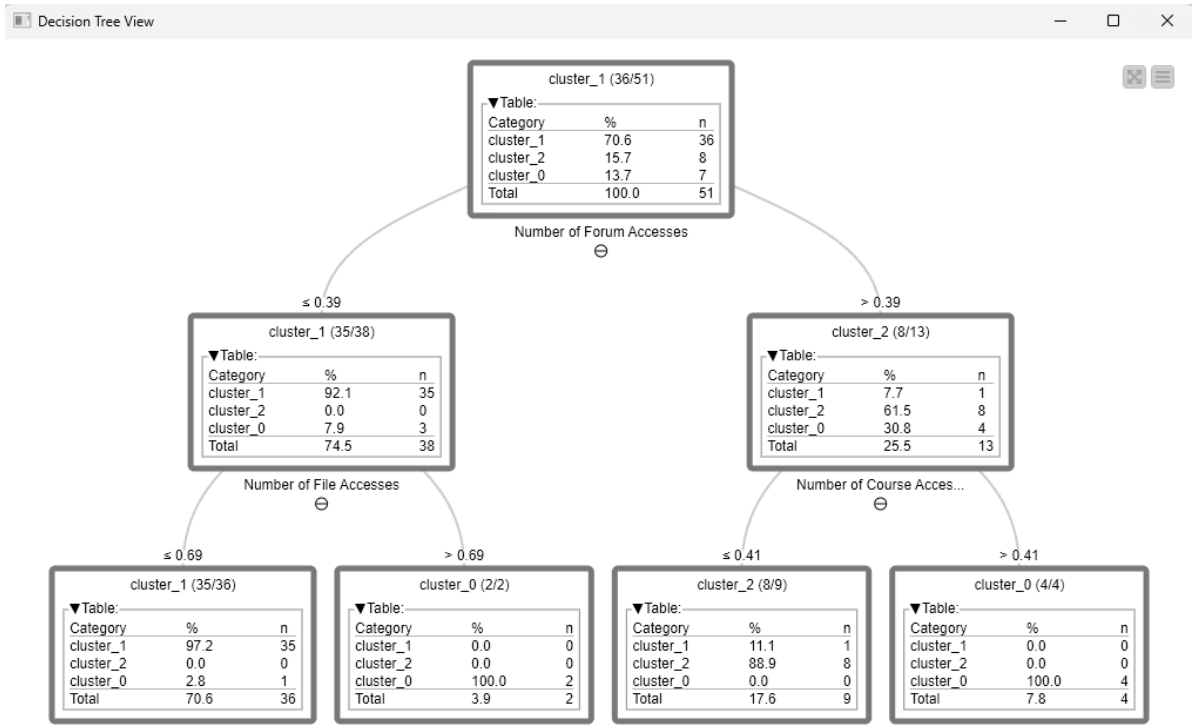


Figure 5. Structure of the decision tree based on student activity clusters.

The visualization of clustering results highlights a clear trend where more active students, particularly those in cluster_0, tend to achieve better academic outcomes, whereas lower engagement, more typical of cluster_2, may correspond to less favourable results. The results of clustering suggest that while moderate activity levels can result in good academic outcomes, students who demonstrate higher levels of activity completion are more likely to obtain better academic outcomes. This reinforces the importance of sustained interaction with the learning platform in influencing students' performance.

The decision tree structure demonstrates that variables related to content engagement, (specifically forum, file, and course accesses) serve as effective discriminators for student grouping. The resulting model provides a transparent interpretation of behavioural patterns that define each cluster and offers insight into how specific types of interaction within the learning environment contribute to distinct student profiles.

A process model illustrating students' engagement during the course is shown in Figure 6. Based on the data, 51 cases were recognized by the process mining tool, performing 52 distinct activities, which occurred a total of 11.9K times. The students had an average of 233 activities (meaning that they interacted with the digital learning platform 233 times) varying from 61 to 896. The median for throughput time from 'process start' to 'process end' is 116 days (approximately the duration of one semester in a higher education institution). Among 10 most frequently executed activities were: Checking course information (in 100% of cases), View File (in 100% of cases), Update data on the completion of e-course activities in the System (in 96% of cases), View Assignment Project Documentation (in 92% of cases), View assignments guidelines Pages (in 90% of cases), and View grade outcomes in Student report (in 88% of cases). The frequency of process activities illustrates students' dedication to check available course content, ensure expected deliverables in the form of assignments made in accordance with assignment guidelines, and their interest in grade outcomes.

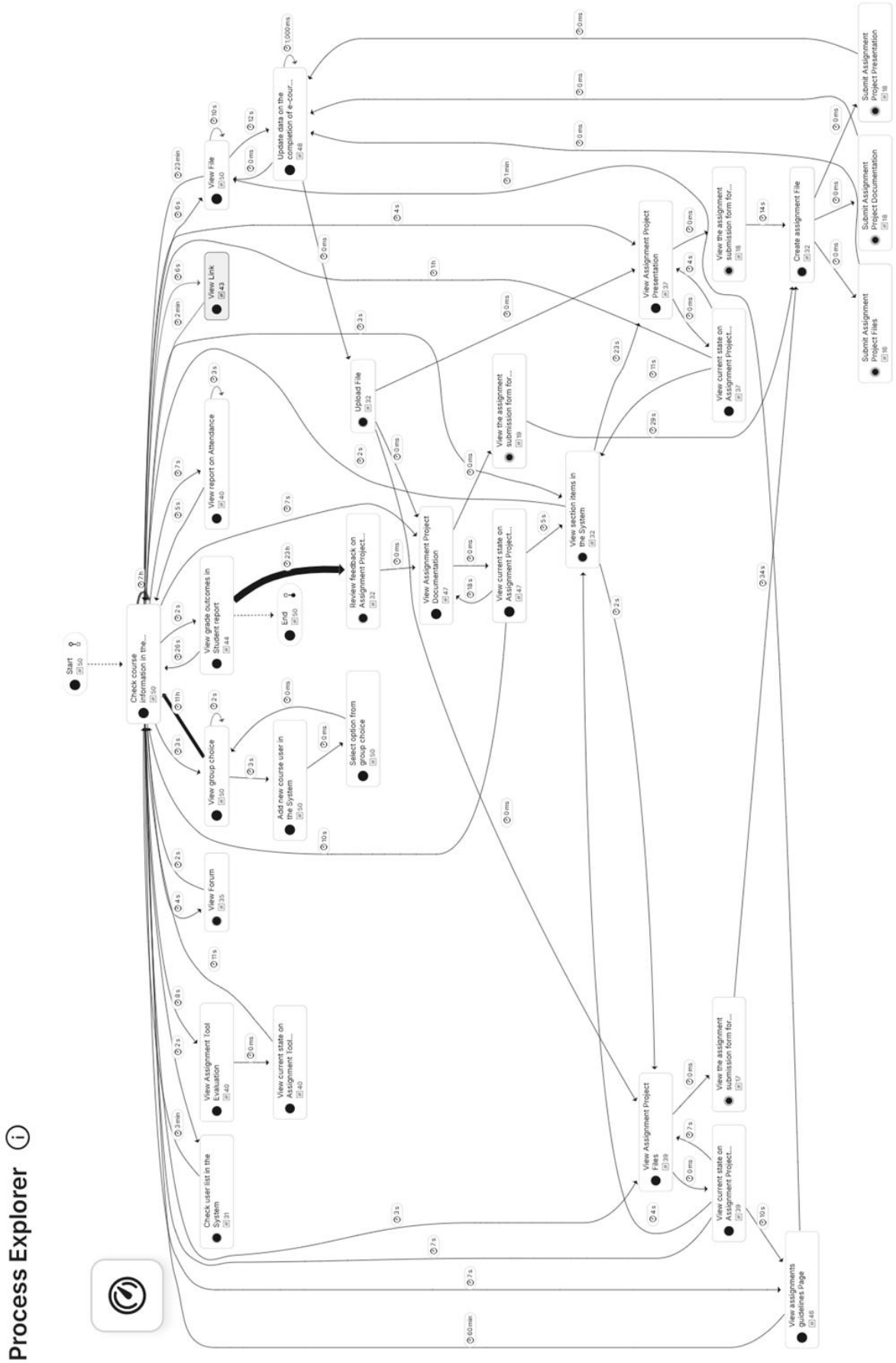


Figure 6. Process model illustrating students' engagement during a university course.

DISCUSSION AND CONCLUSION

The motivation for this article was to investigate student behaviour in using a digital learning platform data and the combination of data and process mining methods to see among other if any exceptional scenarios occurred in relation to previous similar studies, having in mind the disruptive power of technologies in teaching and learning. Combining multiple techniques can lead to more interpretable and actionable insights needed to design and/or implement features or content related improvements in running the course. The value of data and process-oriented analysis in comprehending learning paths of students is relevant to ensure continuous engagement and focus on deliverables in terms of assignment and student projects necessary to complete the course. The initial motive for this research was to investigate if any exceptional scenarios have been recorded based on real-data for one elective course at a higher education institution. Having in mind the rise of gen AI tools as one of the leading useful but also disruptive technologies in teaching and learning, a limitation of this study is that the results were used only at the end of the semester to validate the overall course data. The results showed that no exceptional scenarios occurred in the student behaviour or in their typical learning paths, and second, that the steady and robust design of the course content management applied in similar courses are in line with research results from previous 2020 by Križanić [21] in terms of student engagement and teaching activities.

The findings of this study are consistent with those reported by Križanić [21], who also applied cluster analysis and decision tree techniques to educational log data in order to investigate student behaviour in an e-learning environment. In both studies, clustering enabled the identification of groups of students with distinct engagement patterns, and the decision tree analysis provided interpretable rules that clarified the key behavioural attributes differentiating these groups. Križanić [21] found that variables such as the frequency of accessing learning materials and participation in interactive components were critical in predicting student success, which aligns with the present study's identification of file, forum, and course accesses as the most discriminative features for student grouping. Moreover, both studies highlight the value of combining clustering and decision trees to reveal actionable insights into student engagement and performance. The comparability of results suggests that these techniques are robust in different educational contexts and support their utility for informing course design and personalized learning interventions.

The outcomes highlight a clear relationship between student engagement patterns and academic performance, with higher levels of interaction within the learning platform correlating with better outcomes. In this study, the application of clustering enabled the identification of distinct student groups based on activity levels, while the decision tree structure further supported the interpretation of behavioural patterns within each cluster. These results reinforce the potential of educational process mining and data mining in common to provide deeper insights into student behaviour and to inform the development of personalized learning support strategies based on clearly observed patterns of interaction in digital learning environments. The process model generated based on data on students' interaction with the digital learning platform used in one elective course at a higher education institution, illustrates the diversity of engagement in choosing learning paths through the course (from variations in activity frequency) while retaining the focus on the deliverables expected at the end of the course (submitted and evaluated assignments). Although the analysis did not show exceptional scenarios occurring in the course in terms of deviations in behaviour with the digital learning platform in relation to similar teaching and learning paradigms provided by the same teachers, it can be deducted that more interactive features combined with new technologies would be useful in providing more personalized learning paths as well as those interactions need to be measured continuously to acquire the desired levels of engagement.

This study has several limitations that should be acknowledged. First, the analysis was conducted on data from a single elective course at one higher education institution, which may limit the generalizability of the findings to other contexts or disciplines. Second, only anonymized data recorded within the learning management system were analyzed, without integrating additional contextual or demographic information that could further explain observed behavioural patterns. Third, the data and process mining techniques were applied retrospectively, with findings validated at the end of the semester, thus limiting the potential for real-time interventions during the course.

In further research more in-depth analysis is planned for designing data-based tailored interventions that could be developed and implemented in course management during the semester which could lead towards personalized learning paths. The analysis could be focused on identifying the least-used content from course content, offering to introduce more interactive features of the content which could be combined with new technologies (e.g., gen AI tools). Investigating the effects of introducing new interactive technologies (such as generative AI tools) and measuring their impact on student engagement and learning outcomes could provide valuable insights. Real-time or formative application of process and data mining methods during the course delivery may also enable tailored interventions and more personalized learning support.

REFERENCES

[1] Bey, A. and Champagnat, R.: *Analyzing Student Programming Paths using Clustering and Process Mining*.
In: *Proceedings of the International Conference on Computer Supported Education*, CSEDU. pp.76-84, 2022,
http://dx.doi.org/10.5220/0011077300003182,

[2] Boztaş, G.D.; Berigel, M. and Altınay, F.: *A bibliometric analysis of Educational Data Mining studies in global perspective*.
Education and Information Technologies **29**(7), 8961-8985, 2024,
http://dx.doi.org/10.1007/s10639-023-12170-0,

[3] Chen, A.; Xiang, M.; Zhou, J.; Gašević, D. and Fan, Y.: *Unpacking help-seeking process through multimodal learning analytics: A comparative study of ChatGPT vs Human expert*.
Computers and Education **226**, No. 105198. 2025,
http://dx.doi.org/10.1016/j.compedu.2024.105198,

[4] Dol, S.M. and Jawandhiya, P.M. *Review of EDM for Analyzing the Performance of Students in Educational Setting*.
In: *6th International Conference on Computing, Communication, Control and Automation*, ICCUBEA 2022. 2022,
http://dx.doi.org/10.1109/iccubea54992.2022.10010714,

[5] Dol, S.M. and Jawandhiya, P.M.: *Systematic Review and Analysis of EDM for Predicting the Academic Performance of Students*.
Journal of The Institution of Engineers (India): Series B **105**(4), 1021-1071, 2024,
http://dx.doi.org/10.1007/s40031-024-00998-0,

[6] dos Santos, C.F.; Loures, E.D.F.R. and Santos, E.A.P.: *A smart framework to perform a criticality analysis in industrial maintenance using combined MCDM methods and process mining techniques*.
International Journal of Advanced Manufacturing Technology **136**(9), 3971-3987, 2025,
http://dx.doi.org/10.1007/s00170-024-13193-8,

[7] Feng, G. and Chen, H.: *Educational process mining: A study using a public educational data set from a machine learning repository*.
Education and Information Technologies **30**, 8187-8214, 2024,
http://dx.doi.org/10.1007/s10639-024-13130-y,

[8] Ghazal, M.A.; Ibrahim, O. and Salama, M.A.: *Educational process mining: A systematic literature review*.
In: *Proceedings of the 2017 European Conference on Electrical Engineering and Computer Science* EECS 2017. pp.198-203, 2017,
http://dx.doi.org/10.1109/eecs.2017.45,

[9] Grigorova, K.; Malysheva, E. and Bobrovskiy, S.: *Application of Data Mining and Process Mining approaches for improving e-Learning Processes*.
In: *CEUR Workshop Proceedings*. 115-121, 2017,
http://dx.doi.org/10.18287/1613-0073-2017-1903-115-121,

[10] Hicheur Cairns, A.; Gueni, B.; Hafdi, H.; Joubert, C. and Khelifa, N.: *Towards a distributed computation platform tailored for educational process discovery and analysis*.
International Conference on Protocol Engineering, ICPE 2015 & International Conference on New Technologies of Distributed Systems, NTDS 2015 - Proceedings. No. 7293494, 2015,
http://dx.doi.org/10.1109/notere.2015.7293494.

[11] Intayoad, W.; Kamyod, C. and Temdee, P.: *Process mining application for discovering student learning paths*.
In: *3rd International Conference on Digital Arts, Media and Technology*, ICDAMT 2018. 220-224, 2018,
http://dx.doi.org/10.1109/icdamt.2018.8376527,

[12] Juhaňák, L.; Zounek, J. and Rohlíková, L.: *Using process mining to analyze students' quiz-taking behavior patterns in a learning management system*.
Computers in Human Behavior **92**, 496-506, 2017,
http://dx.doi.org/10.1016/j.chb.2017.12.015,

[13] Levin, N.A.: *Process Mining Combined with Expert Feature Engineering to Predict Efficient Use of Time on High-Stakes Assessments*.
Journal of Educational Data Mining **13**(2), 1-15, 2021,
http://dx.doi.org/10.5281/zenodo.5275310,

[14] Ramaswami, G.; Susnjak, T.; Mathrani, A.; Lim, J. and Garcia, P.: *Using educational data mining techniques to increase the prediction accuracy of student academic performance*.
Information and Learning Science **120**(7-8), 451-467, 2019,
http://dx.doi.org/10.1108/ils-03-2019-0017,

[15] Romero, C.; Cerezo, R.; Bogarín, A. and Sánchez-Santillán, M.: *Educational process mining: A tutorial and case study using moodle data sets*.
In: ElAtia, S.; Ipperciel, D. and Zaïane, O.R., eds.: *Data Mining And Learning Analytics: Applications in Educational Research*. Wiley, pp.1-28, 2016,
http://dx.doi.org/10.1002/9781118998205.ch1,

[16] Salazar-Fernandez, J.P.; Munoz-Gama, J. and Sepúlveda, M.: *Implications of losing a need- and merit-based scholarship on the educational trajectory: a curricular analytics approach*.
Higher Education **89**(2), 441-464, 2025,
http://dx.doi.org/10.1007/s10734-024-01230-0,

[17] Simic, D.; Leible, S.; Schmitz, D.; Gücük, G.-L. and Kučević, E.: *Enhancing Project-based Learning through Data-driven Analysis and Visualisation: A Case Study*.
https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1139&context=acis2023,

[18] Wibawa, B.; Siregar, J.S.; Asrorie, D.A. and Syakdiyah, H.: *Learning analytic and educational data mining for learning science and technology*.
AIP Conference Proceedings **2331**, No. 060001. 2021,
http://dx.doi.org/10.1063/5.0041844,

[19] Yunita, A.; Santoso, H.B. and Hasibuan, Z.A.: *Research Review on Big Data Usage for Learning Analytics and Educational Data Mining: A Way Forward to Develop an Intelligent Automation System*.
Journal of Physics: Conference Series **1898**(1), No. 012044. 2021,
http://dx.doi.org/10.1088/1742-6596/1898/1/012044,

[20] Kitchenham, B.; Pearl Brereton, O.; Budgen, D.; Bailey, J. and Linkman, S.: *Systematic literature reviews in software engineering - A systematic literature review*. Information and Software Technology **51**(1), 7-15, 2009, http://dx.doi.org/10.1016/j.infsof.2008.09.009,

[21] Križanić, S.: *Educational data mining using cluster analysis and decision tree technique: A case study*. International Journal of Engineering Business Management, No. 12, 2020, http://dx.doi.org/10.1177/1847979020908675,

[22] Celonis: *When processes work, AI works*. https://www.celonis.com.