Discover Education

Research

Opportunities of machine learning algorithms for education

Olga Ovtšarenko[1,2]

Received: 9 September 2024 / Accepted: 21 October 2024
Published online: 04 November 2024

Abstract
Machine learning (ML) methods are among the most promising technologies with wide-ranging research opportunities, particularly in the field of education, where they can be used to enhance student learning outcomes. This study explores the potential of machine learning algorithms to build and train models using log data from the "3D Modeling" e-course on the Moodle platform at TTK University of Applied Sciences, Tallinn, Estonia. By predicting trends, identifying patterns, and optimizing resource allocation, machine learning can improve the efficiency of e-learning and provide students with tailored recommendations for acquiring relevant knowledge and skills. The results of the study show that machine learning algorithms can be used to process the available e-course log data, using the clickstream of e-course resources and for their automated processing. The results suggest potential applications in personalized course recommendations, prediction and dropout prevention strategies, resulting in a more effective and personalized educational experience. Future research will focus on improving models of available registration data, exploring and using advanced machine learning techniques to improve the accuracy and usefulness of predictions, and providing faster recommendations to help students navigate their studies more effectively.

Keywords Moodle · e-learning · Logging data · Data processing · Machine learning algorithms · Features engineering · Model training

1 Introduction

The rapid advancements in data-related technologies are reshaping education, fostering optimistic views about how data can inform educational reform and drive positive outcomes for both institutions and students [1]. The growth of technology in education has revolutionized learning by offering personalized content through e-learning platforms. To support (ability to adapt), collecting and processing student data for automated learning pathway creation is critical. This process offers several advantages:

- Collected data about students, such as their performance, preferences, learning style, and pace, can be used to design personalized educational programs that allow each student to progress at their own speed and skill level,
- Data processing can identify areas where students struggle, enabling educators to provide supplementary materials or activities to close knowledge gaps and improve learning efficiency,
- Tracking student performance data allows for real-time adjustments to instructional materials and methods, better addressing student needs,

✉ Olga Ovtšarenko, olga.ovtsarenko@tktk.ee | [1]TTK University of Applied Sciences, Tallinn, Estonia. [2]Vilnius Gediminas Technical University, Vilnius, Lithuania.

- Data analysis can help identify students at risk of falling behind and proactively offer additional resources or support, preventing learning difficulties,
- Analyzing the use of educational resources enables more efficient allocation of those resources, focusing on the technologies and methods that work best for specific student groups (Fig. 1).

While the availability of vast amounts of information is a boon, it also presents challenges. The overwhelming volume of data makes searching, filtering, and selecting relevant information difficult [2]. Adaptive e-learning systems, which track student progress to personalize content delivery, are a solution to this problem. These systems adjust content and formatting to maximize cognitive engagement and streamline students' navigation through the information overload. Investigating which elements of learning should be adapted, and to what extent, is a core focus of adaptive e-learning systems.

The goal of an adaptive e-learning recommendation system is to improve student outcomes. Secondary objectives include (a) relevance (learning outcomes must be pertinent to avoid becoming obsolete), (b) novelty (introducing innovative or unused content), (c) uncertainty (unexpected but relevant learning outcomes), and (d) diversity (varied learning outcomes to broaden subject coverage). Additionally, certain criteria are required for effective recommendations: (a) transparency (explaining the recommendation), (b) confidence (supporting student engagement with learning outcomes), (c) satisfaction (enhancing student outcomes), (d) persuasion (motivating students to follow recommendations), (e) efficiency (assisting in decision-making), and (f) effectiveness (helping students make informed decisions) [3].

This study leverages data from the Moodle platform at TTK University of Applied Sciences (TTK UAS), with the following key characteristics:

- Moodle is one instance of the open source learning management system Moodle.
- Moodle offers teachers the ability to create and manage courses and students the opportunity to participate in the same courses.
- Moodle users are natural persons who have a Moodle account.
- Users can only use Moodle if they agree to all terms of use (following the rules of Protection of personal data [4]).
- The use of Moodle mainly means the use of Moodle's possibilities to participate in the educational activities/ to carry out a study.

The structure of this paper is as follows: Sect. 1 presents the literature review, while Sect. 2 discusses the methodology and the processing of available data on the Moodle platform. Section 3 outlines how machine learning algorithms are applied for data processing—filtering, visualization, and clustering. Section 4 explains the training of the log data model and its improvement. Section 5 presents the results analysis, and Sect. 6 concludes the study with a discussion of future work.

2 Literature review

Rapid advances in technology have transformed learning, empowering it with online resources and making it the dominant mode of knowledge delivery. Improving its effectiveness by using data analytics to predict trends and optimize resource allocation to improve student learning outcomes and provide targeted recommendations for acquiring relevant knowledge and skills. This literature review summarizes studies on trend prediction and resource optimization in e-learning—examining their impact on learning outcomes.

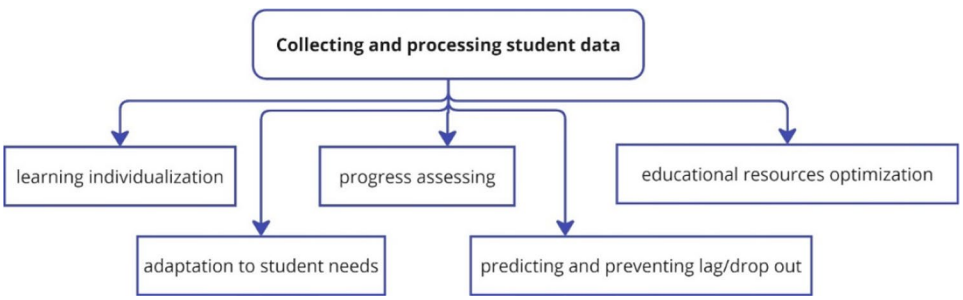


Fig. 1 Reasons for collecting and processing student data

Hussain et al. [5] used machine learning algorithms to predict student performance based on previous learning behavior, suggesting that personalized learning paths can be created to improve outcomes. Such predictive models help identify students at risk of failure or disengagement, allowing educators to intervene early.

Using data mining techniques, educators can discover hidden patterns that reveal how students engage with learning materials, manage their time, and progress through the curriculum. By analyzing clickstream data, time spent on different modules, and assessment results, researchers can determine which resources or teaching methods are most effective for specific groups of learners. Romero and Ventura [6] further discuss how pattern identification can be used to tailor content delivery to individual learners' needs, supporting differentiated learning and promoting better engagement.

Worsley and Blickstein [7] highlight that pattern identification can provide insights into learners' cognitive processes, allowing educators to tailor their teaching methods accordingly. Identifying patterns in learners' errors can help develop more effective feedback mechanisms, which are critical to improving learning outcomes in e-learning environments.

Efficient resource allocation ensures that educational institutions can provide learners with the appropriate support, technology, and content in a cost-effective manner. Tsai et al. [8] suggest optimizing learning management systems by analyzing learner usage patterns and feedback to allocate resources where they are needed most. Adaptive learning platforms that personalize content based on learner performance require significant data and computing resources.

Lamas et al. [9] propose that the integration machine learning algorithms into e-learning platforms allows for dynamic allocation of resources. Their research demonstrates that Artificial Intelligence (AI) based systems can recommend learning materials, quizzes, and feedback sessions based on real-time student data, leading to a more efficient and effective learning process.

Wang et al. [10] describes that student who received personalized recommendations based on predictive and analytical tools demonstrated a significant improvement in their academic performance compared to those who followed a traditional, non-personalized learning path.

The effectiveness of these techniques often depends on the quality of the data collected, which may vary across institutions and platforms. Current research must focus on addressing these challenges. The literature demonstrates that data-driven techniques can lead to more personalized, adaptive, and effective learning environments.

To create individual learning paths that match the characteristics and abilities of each student using multiple choice test scores to predict language proficiency levels, Imamah et al. [11] used models of the relationship between students' responses and their latent abilities to improve the prediction of language proficiency levels, using individual learning path dynamics that take into account changes in students' abilities in each module. Machine learning algorithms, analyzing patterns and relationships in learning data (analysis of initial learning log related to study of Hasegawa et al.) [12], can be used to better predict student ability and create more optimal learning paths. The disadvantage of dynamic personalized learning is that iterative learning occurs repeatedly before students reach the passing score, which can often lead to boredom. Therefore, it is necessary to implement dynamic question randomization methods adapted to student ability and a cross-testing system that will provide a more effective assessment of student performance improvements.

A literature review shows that data-driven methods can lead to more personalized, adaptive, and effective learning environments to improve student education. And using available e-learning platform log data to track student activity and learning resource usage can be a useful and effective tool to improve learning results.

3 Methodology

Extracting value from data faces several practical challenges, including discovering relevant data, extracting information from primary sources, identifying relationships between datasets, and resolving inconsistencies in representation [13].

Figure 2 illustrates the planned stages of data use: 1—registration of student data, including recording and coding; 2—use of encoded data and filtering; 3—possible grouping students based on shared characteristics to perform specific level tasks.

Data overlay enables the simultaneous viewing of multiple logging sessions, and the collected data can be easily exported to third-party software for detailed analysis.

Understanding the ethical considerations associated with data recording and automated e-learning project creation is crucial. Privacy and data security must be prioritized to protect students' personal information. E-learning platforms and educational institutions should implement robust policies and procedures to handle data responsibly and transparently.

Since registration data alone is insufficient for generating personalized learning trajectories, the author plans to collect additional data through student questionnaires, ensuring student consent for using the information. Feedback from

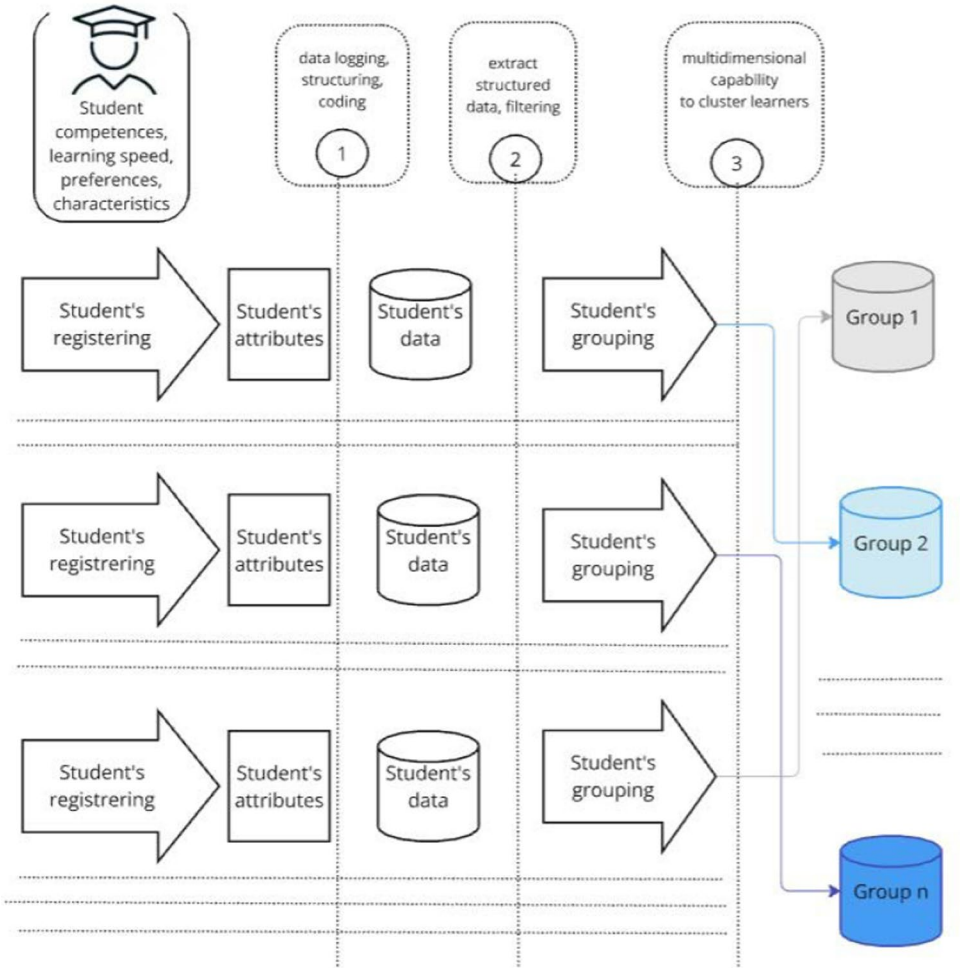


Fig. 2 Data processing stages

students regarding their usage of e-course resources will also be collected to further aggregate attributes that influence grouping by shared characteristics. The questionnaire design will adhere to the General Data Protection Regulation (GDPR), providing essential features for initial grouping (Fig. 2).

Data preparation is critical, involving key functions such as profiling, matching, mapping, format transformation, and data repair (Fig. 3). Current trends in data preparation emphasize convergence—combining different datasets and managing separate datasets—and automation to guide user actions [14].

This study aims to employ deep learning algorithms to identify learner attributes based on available registration data on the Moodle platform, ultimately facilitating the automatic creation of personalized learning paths for individual

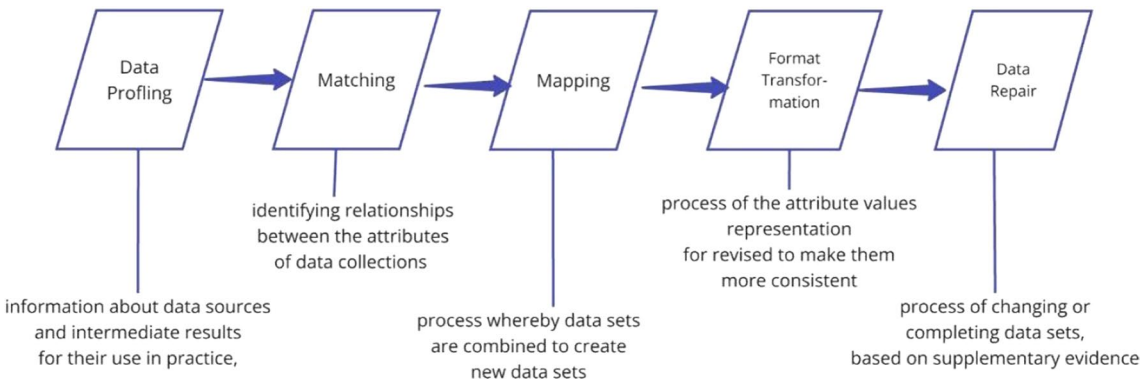


Fig. 3 Structure of key functions of data preparation

learners or small groups. Personalized e-learning experiences are tailored to meet the specific needs of each learner. By analyzing data collected through registration, e-learning platforms can automatically generate these pathways, considering academic performance and learning progress [15].

3.1 Moodle E-course available log data

Data logging, the systematic collection and recording of information, has become widespread in education [16]. It provides educators with valuable data to enhance teaching quality. The demand for personalized learning paths has further driven the development of automated systems that tailor educational programs to individual student profiles.

Data logging captures, stores, and displays datasets for analyzing activity, identifying trends, and predict future events. The primary goal data logging is to analyze user interaction with the system, identify trends. Continuous monitoring and analysis of data provides information about current processes, which allows for informed forecasts and optimizations. Most data-logging processes are automated through intelligent applications like artificial intelligence, machine learning, or robotic process automation [17].

Log data in e-learning systems captures learner behaviors and experiences. This data includes metrics like browsed content types, time spent on various pages, number of clicks, number of e-assessment tasks, and interactions among users (student–student or student-instructor) [18].

Log data comprises a chronological sequence of single or multi-line events generated by applications, capturing specific system states. Log events typically include a creation timestamp, logging levels (e.g., INFO or ERROR), and process IDs that connect related event sequences [19].

In this study, data on learning resource usage was collected from the Moodle platform at TTK University of Applied Sciences (TTK UAS), Estonia. Log data from a group of twenty students gathered at the end of the Spring 2024 semester. The "3D modelling" e-course is a 3 ECTS subject offered to students from all specialties at TTK UAS, both part-time and full-time. The proposed teaching method for this course includes hybrid and independent study of educational materials and practical exercises, utilizing interactive resources to facilitate successful training for students in diverse fields.

The course spans eight weeks, with four academic hours each week, covering various modeling topics to develop spatial imagination and computer skills for 3D modeling, while also fostering engineering knowledge in reading and drawing.

Statistics in Moodle [20]:

- The statistics graphs and tables show how many hits there have been on various parts of the site during various periods.
- Logs in Moodle are activity reports.
- Logs are available at site level and course level.

This study explored the capabilities of obtaining Moodle statistics and using existing data to aggregate new attributes, online behavior, and characteristic groupings. Figure 4 presents data extracted from TTK UAS Moodle statistics. Moodle e-course filters were applied to extract data within the planned learning period (January 29 to April 15, 2024). The dataset contained no missing values and consisted of 7309 rows and 9 columns detailing educational resource usage, including Time, User full name, Affected user, Event content, Component, Event name, Description, Origin, and IP address.

Data modeling involves using mathematical models and statistical assumptions to:

- Create sample data.
- Establish mathematical relationships between random and non-random variables.

	Time	User full name	Affected user	Event context	Component	Event name	Description	Origin	IP address
2	15/04/24, 11:37:03	A.....B.....	-	Task: 3.homework - for checking	Task	The status of the submission has been viewed.	The user with id '26861' has viewed the submission status page for the assignment with course module id '233243'.	web	1....1....6...3.
3	15/04/24, 11:37:03	A.....B.....	A.....B.....	Task: 3.homework - for checking	Task	Feedback has been viewed	The user with id '26861' viewed the feedback for the user with id '26861' for the assignment with course module id '233243'.	web	1....1....6...3.
4	15/04/24, 11:37:03	A.....B.....	-	Task: 3.homework - for checking	Task	Course module has been viewed	The user with id '26861' viewed the 'assign' activity with course module id '233243'.	web	1....1....6...3.

Fig. 4 Students log data journal

- Apply statistical methods to describe and analyze relationships between variables in the system.
- Evaluate uncertainty within the system under study.
- Predict outcomes.
- Generate real-world forecasts.

3.2 Data preprocessing

Data cleaning is the process of identifying and correcting inaccurate data [21]. Before further analysis, it is essential to thoroughly clean the data, which ultimately enhances data quality.

For the specified period, extracted data underwent cleaning—teacher logs were deleted, and student personal data was replaced to comply with the General Data Protection Regulation (GDPR). The data were subsequently prepared, coded, and aggregated.

The objective of this study's data mining is to analyze the behaviors, priorities, and challenges faced by a specific group of students using educational materials.

However, extracting statistics on the usage of learning resources from the Moodle platform is limited due to minimal registration data and the constraints of the platform's filters for collecting statistics, which do not allow retrieval of data for specific periods (e.g., a school year or a single school day).

Therefore, Excel functions were employed to clean the data, replace values, and obtain additional information during the model preparation phase (Fig. 5).

To prepare the data, the columns in the data table were populated with the following information [22]:

1. timestamp,
2. user_id—student names were replaced with a user registration codes on the Moodle platform from the column "Description" context,
3. tool_id—names of the resource used were replaced with registration codes from the Moodle e-course repository, sourced from the "Description" context,
4. log_count—data (e.g., 26,861 clicks by one user within a specific timeframe) was aggregated to capture interaction frequency,
5. lesson—aggregated data indicating whether resource usage aligned with scheduled lesson times—1 for a match, otherwise 0,

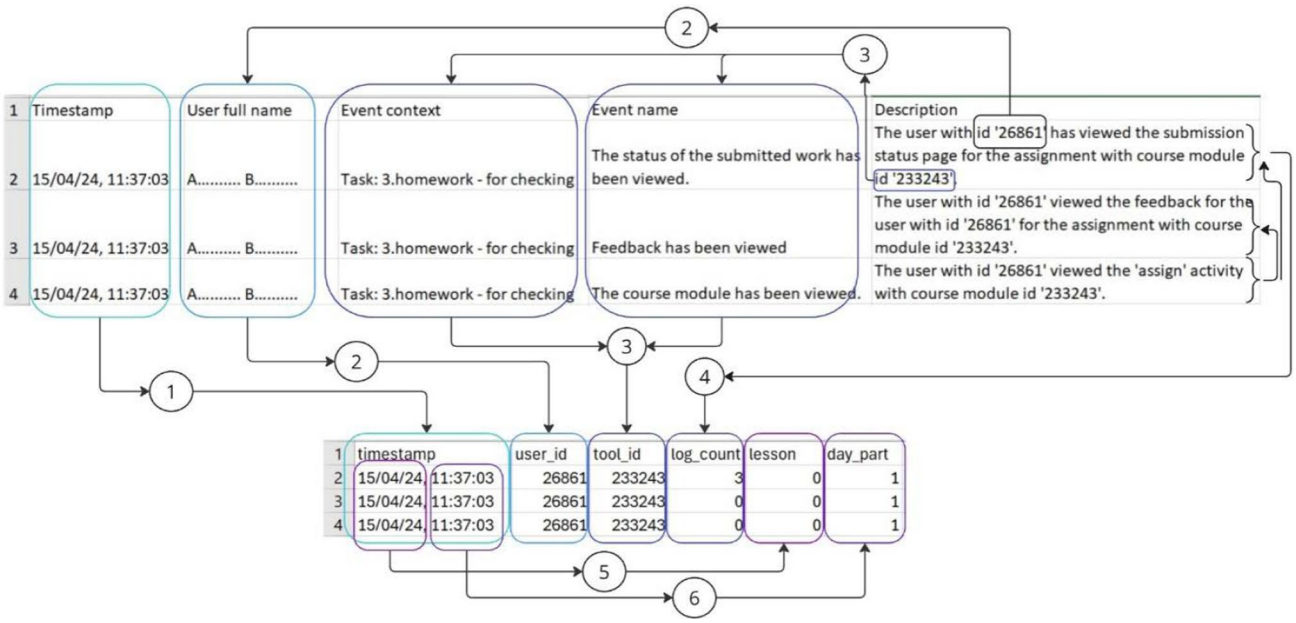


Fig. 5 Data processing

6. day_part—aggregated data categorizing usage by part of the day in four-hour intervals (1 = 08:00–12:00; 2 = 12:00–16:00; 3 = 16:00–20:00; 4 = 20:00–00:00; 5 = 00:00–04:00; 0 = 04:00–08:00).

As a result, the preprocessed data model consisted of 6096 rows and 6 columns instead of the original 7309 rows and 9 columns. This cleaned dataset will be used to apply machine learning algorithms for analyzing usability, training models, identifying patterns, and making predictions.

4 Data processing

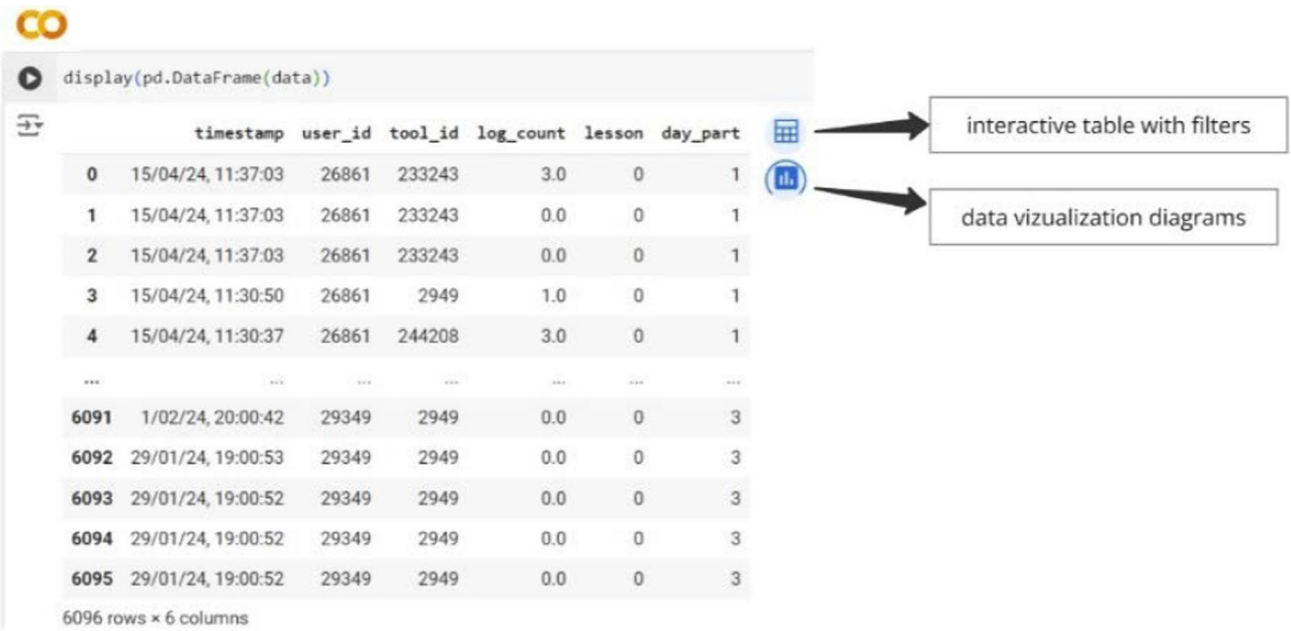
Machine learning employs various algorithms to uncover patterns in diverse data formats such as documents, images, and audio. A multi-level learning process enhances performance; the initial level identifies primary characteristics, while subsequent levels add complexity, thereby improving identification accuracy. The effectiveness of the model is assessed by comparing predicted values against actual outcomes [23]. For data analysis, algorithms from the PyTorch library were implemented within interactive Colab Notebooks using Python.

4.1 Machine learning algorithms opportunities for data filtering and visualization

To utilize data tables (.xlsx or.csv) within the interactive Colab Notebook, the files must be accessible via a Google Drive link or imported into the notebook's storage for the session. Figure 6 illustrates basic operations on the data table with buttons for interactive table use (Fig. 7) and data visualization (Fig. 8).

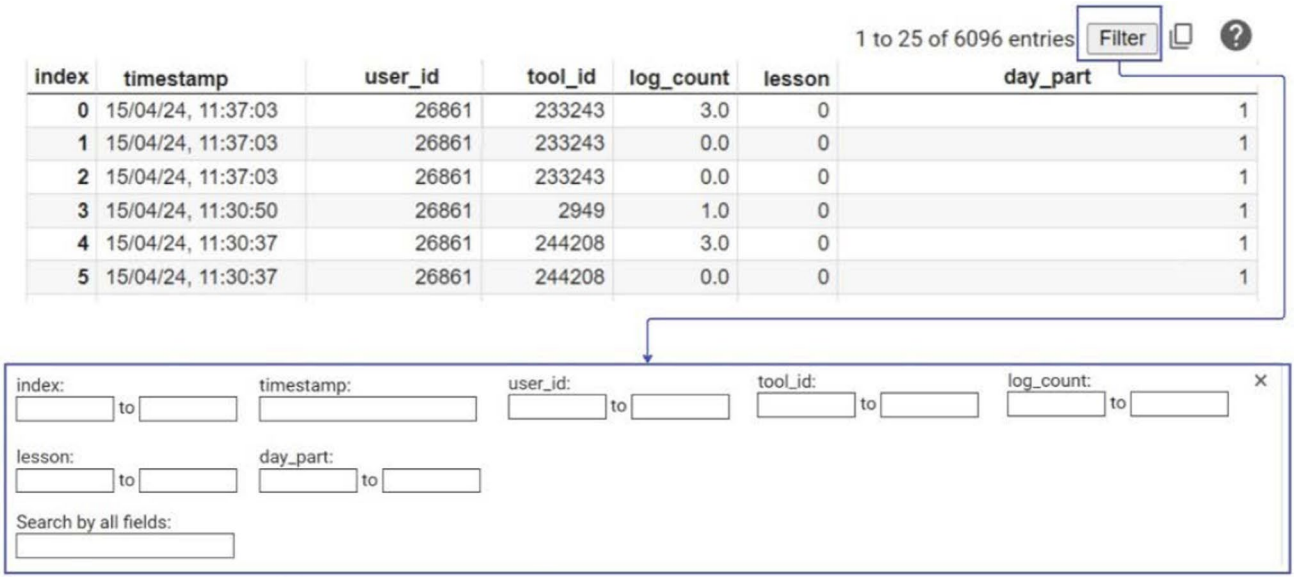
When extracting data from Moodle's educational resource usage logs, the inherent limitations of the platform's filters (e.g., selecting records for a single student, an entire group, or a specific day) necessitate the implementation of column-based filtering. This approach allows for precise viewing, selection, and analysis of the data [24].

Visual data representation through charts, graphs, and histograms facilitates immediate comparison and recognition of patterns [23]. Visual representation of data highlights the main facts and relationships between the data. It is possible to compare data using charts to analyze data in various forms of information. Also forecasting capability—schematic and graphical representation of information has past patterns, which helps in analyzing and forecasting various strategies for the future. Key advantages of data visualization include:



	timestamp	user_id	tool_id	log_count	lesson	day_part
0	15/04/24, 11:37:03	26861	233243	3.0	0	1
1	15/04/24, 11:37:03	26861	233243	0.0	0	1
2	15/04/24, 11:37:03	26861	233243	0.0	0	1
3	15/04/24, 11:30:50	26861	2949	1.0	0	1
4	15/04/24, 11:30:37	26861	244208	3.0	0	1
...	...	...	...	...	...	...
6091	1/02/24, 20:00:42	29349	2949	0.0	0	3
6092	29/01/24, 19:00:53	29349	2949	0.0	0	3
6093	29/01/24, 19:00:52	29349	2949	0.0	0	3
6094	29/01/24, 19:00:52	29349	2949	0.0	0	3
6095	29/01/24, 19:00:52	29349	2949	0.0	0	3

Fig. 6 Data processing



index	timestamp	user_id	tool_id	log_count	lesson	day_part
0	15/04/24, 11:37:03	26861	233243	3.0	0	1
1	15/04/24, 11:37:03	26861	233243	0.0	0	1
2	15/04/24, 11:37:03	26861	233243	0.0	0	1
3	15/04/24, 11:30:50	26861	2949	1.0	0	1
4	15/04/24, 11:30:37	26861	244208	3.0	0	1
5	15/04/24, 11:30:37	26861	244208	0.0	0	1

Fig. 7 Data filtering

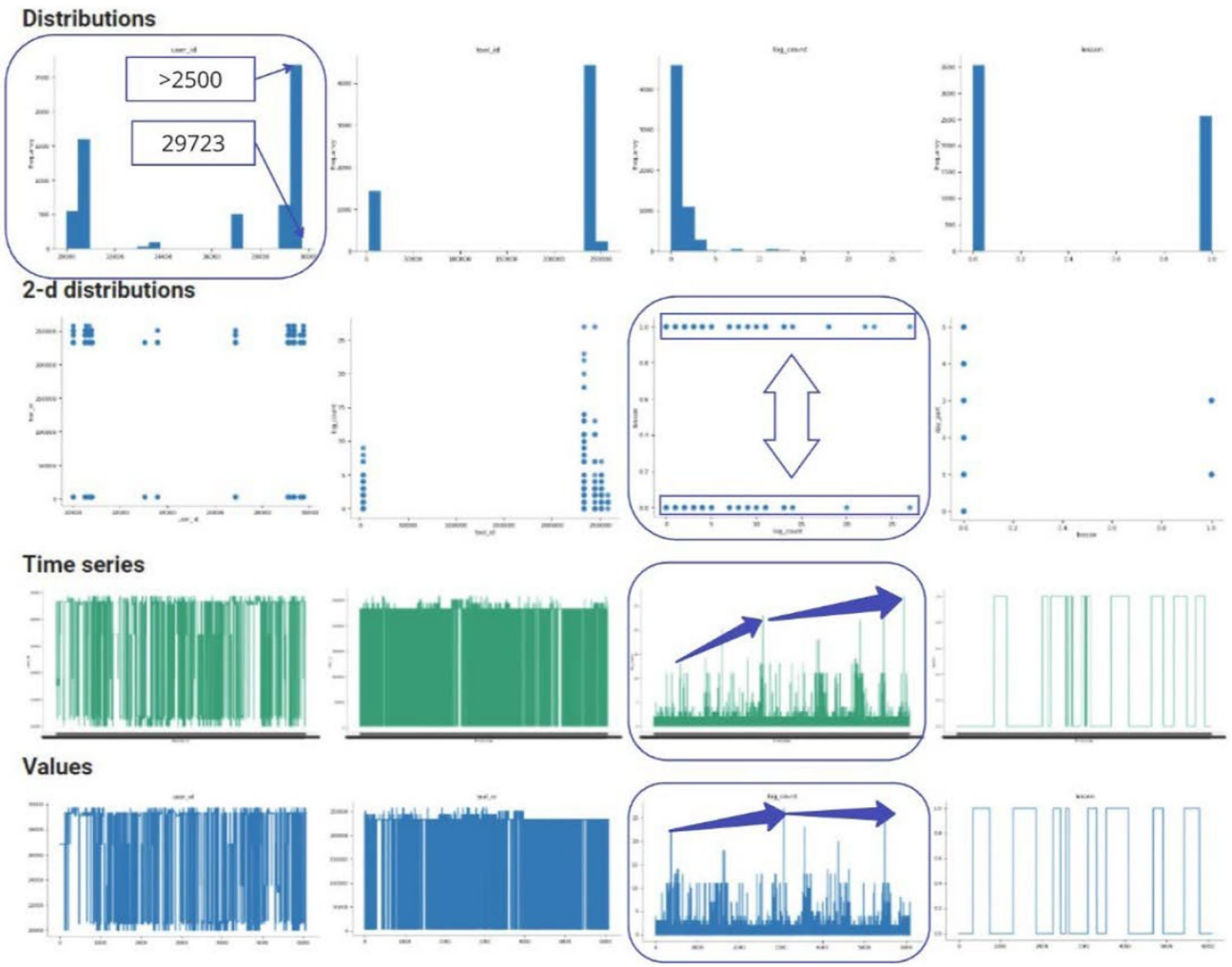


Fig. 8 Data visualization diagrams

- Diagrams present complex data in a simplified form. Hence, the facts presented in the form of diagrams can be understood quickly.
- Diagrams make comparisons easier.
- Diagrams are used to present massive amounts of complex data in a simplified and understandable format.
- The image created in the mind by diagrams lasts much longer compared to images created by numbers presented in tabular form.
- Diagrams provide more information. Charts not only display the characteristics of data, but also reveal hidden facts and relationships that cannot be obtained from classified and tabular data [24].

Data visualizations (Fig. 8) [22] are categorized into four groups: Distributions, 2D Distributions, Time Series, and Values. Notably, the Distributions bar charts provide insights into student activity and the most utilized resources, such as the Frequency of usage against user_id and tool_id. The data suggest that first-year students, who registered later, were highly active users of the e-course resources, particularly the H5P interactive book featuring brief videos (1–3 min.) on furniture modeling.

2D Distributions facilitate the examination of two parameter relationships, aiding in data grouping and identifying correlations. Point diagrams plotting tool_id against user_id and log_count against user_id reveal student preferences and online behavior, allowing for further student attribute aggregation. The lesson versus log_count diagram highlights significant independent study outside scheduled lessons, indicating a proactive approach to learning.

Time Series histograms and bar charts present data trends over the academic period, showcasing student engagement and resource usage dynamics. The log_count versus timestamp graph demonstrates a consistent increase in student activity throughout the semester.

Value charts chronologically review parameter values, exhibiting similarities with the log_count versus timestamp diagram, albeit with variations in vertical values indicating limited change in logs over time.

4.2 Data correlation

Value charts chronologically review parameter values, exhibiting similarities with the log_count versus timestamp diagram, albeit with variations in vertical values indicating limited change in logs over time [25]. Correlation of two variables can be positive—when one increases, the other also increases, or negative—when one goes up, the other goes down.

Correlations are measured by the correlation coefficient and can be categorized as:

Ideal correlation: Perfect correlation occurs when two variables change proportionately, yielding:

Positive correlation (+ 1).

Negative correlation (− 1).

Zero correlation: no relationship exists between variables, leading to no influence from one on the other.

Limited degree of correlation: correlation coefficients fall between + 1 and -1, indicating varying degrees of relationship:

Low (0 to 0.25).

Medium (0.25 to 0.75).

High (0.75 to 1).

Results from deep learning algorithms revealed a weak correlation among the attributes utilized [26] and results diagram is presented on Fig. 9.

Lesson and day_part: − 0.293 (weak negative correlation). As one feature's value increases, the other feature's value tends to decrease slightly. A weak negative correlation may indicate that certain lessons are less likely to be scheduled at certain times of the day.

Lesson and tool_id: 0.14 (very weak positive correlation). The very weak positive correlation suggests that tool selection for lessons is independent of lesson content.

The gained results of correlation alone cannot predict future values, but it can be part of predictive models. Correlation analysis helps identify relationships that can be further explored with regression analysis and other predictive modeling techniques. To determine the interaction between several functions, it is advisable to use additional analysis. It is also necessary to aggregate and explore other functions that may have closer relationships.

Understanding the types of correlations is critical to interpreting data generated in an educational context [27]. Understanding these connections makes it possible to develop effective interventions to help students, tailor the educational experience, and improve learning outcomes.

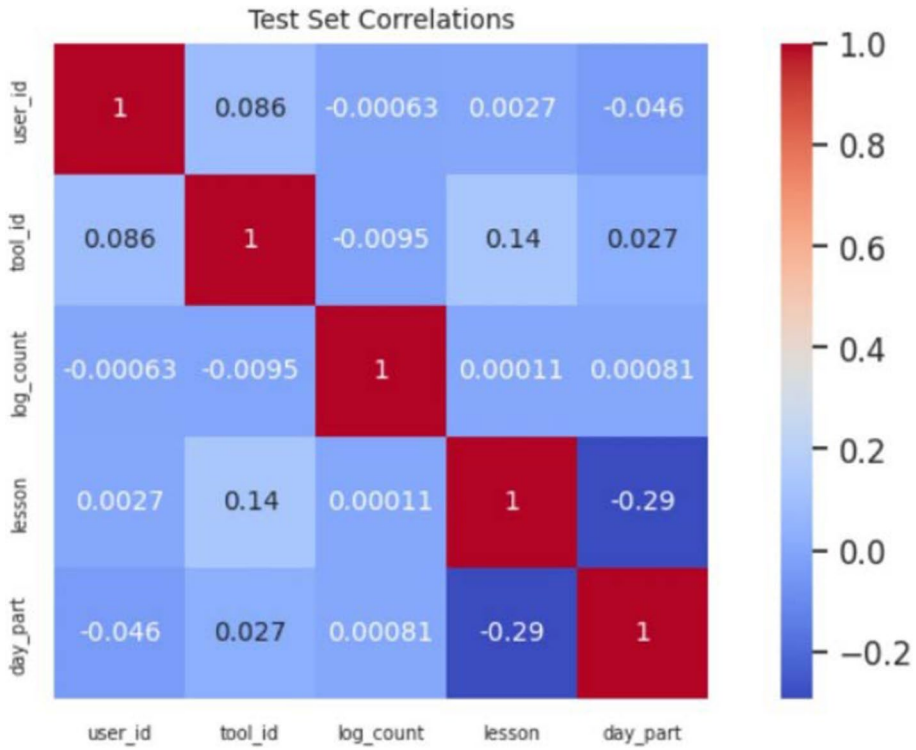


Fig. 9 Features correlation

Interpreting correlations requires careful consideration of context. For example, a positive correlation between technology use in the classroom and student engagement does not necessarily mean that more technology leads to more engagement. Other contributing factors must be considered, such as teaching quality or student motivation, which are not captured by simple correlation analysis.

Correlation does not imply causation. The fact that two variables move together does not mean that one causes the other. This is a common misunderstanding that needs to be considered when making decisions based on correlational data.

Correlation is a principal issue in educational research, providing insight into the complex relationships between different elements of the learning process. By understanding the several types of correlations, it is possible to better appreciate the complex structure of factors that influence the success of the educational process and use it to personalize it.

4.3 Data clustering

Cluster analysis effectively identifies data patterns by categorizing similar objects based on their characteristics [28]. Discovering hidden relationships in data can be done by identifying clusters in the data 'log_count' vs 'user_id' and obtaining information about their underlying structure [29]. There is needed to select the type of clustering that is appropriate for the data we have, select the appropriate clustering method, and interpret the results.

Common clustering methods include hierarchical, partitioned, density-based, and model-based clustering. Each method has unique advantages and limitations based on data type and clustering goals. Comparing results across various algorithms enhances pattern identification and accuracy.

The hierarchical clustering dendrogram [30] illustrates data point merges through horizontal lines, as seen in Fig. 10.

Interpretation of the dendrogram reveals clusters formed at heights of 12,500, 25,000, and 50,000 logs, indicating similar data point proximity within clusters and distinguishing them from others.

Cutting the dendrogram at various heights reveals 3 clusters at 50,000, 4 clusters at 25,000, and 6 clusters at 12,500. The identified cluster counts were further analyzed using the K-means clustering algorithm to ensure appropriate grouping.

The obtained cluster numbers are used in the K-means clustering algorithm to ensure that similar data are clustered together, while dissimilar data are in different clusters.

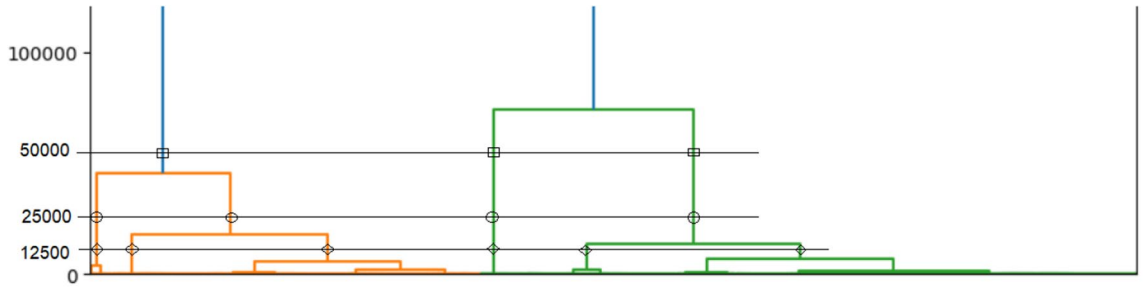


Fig. 10 Dendrogram with cutting levels

K-Means Clustering categorizes unlabeled datasets, allowing for the discovery of groupings without prior training. It operates as a centroid-based algorithm aimed at minimizing the distance between data points and their respective clusters [31].

Initial results from applying K-means with 3, 4, 5, and 6 clusters are summarized in Table 1, utilizing the Silhouette score to evaluate clustering quality [32].

The Silhouette Score quantifies how closely data points align with their clusters compared to others [33].

Ranging from − 1 to 1, a high score signifies strong cluster alignment. The analysis indicated that the optimal number of clusters was 4, yielding a Silhouette score of 0.733. Thus, the subsequent analysis focuses on the four-cluster outcomes displayed in Fig. 11.

Results for 4 Clusters:

The centroids for each cluster, characterized by user_id and log_counts, are as follows:

Cluster 1 Centroid: [29360.5682, 0.966877447]

Cluster 2 Centroid: [20469.0667, 0.969216418]

Table 1 K-means clustering algorithm used for Silhouette score

Clusters number	3	4	5	6
Silhouette score	0.708	**0.733**	0.679	0.621

3rd cluster (5000 logs)

4th cluster (25000 logs)

5th cluster (12500 logs)

6th cluster (12500 logs)

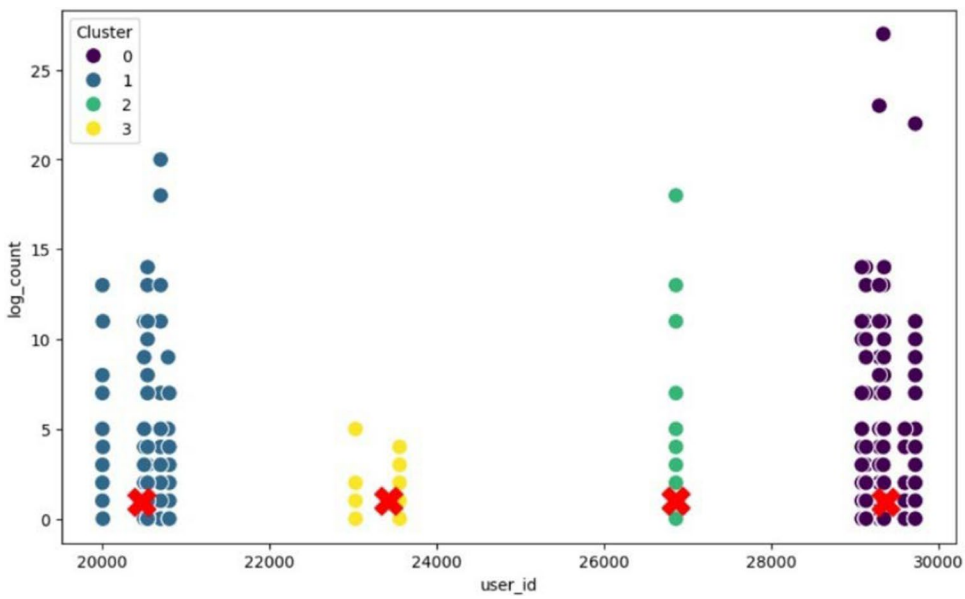


Fig. 11 K-means Clustering with 4 Clusters

Cluster 3 Centroid: [26861.0000, 0.976190476]
Cluster 4 Centroid: [23428.0952, 0.992063492]

The centroids reflect similar log_count values, indicating close similarities among clusters. By centroid distances, infer the relationships and distinctions between the clusters, helping to understand the underlying structure of the data and providing insights for further analysis or decision-making. These distances indicate how far apart the centroids are from each other in the feature space. A lower distance indicates that clusters are closer to each other, whereas a higher distance indicates more separation. The analysis of centroid distances reveals the relationship between clusters, guiding further exploration:

Centroid Distances:
Cluster 1 and Cluster 2: 8891.5015 (well-separated)
Cluster 1 and Cluster 3: 2499.5682 (relatively close)
Cluster 1 and Cluster 4: 5932.4729 (moderately separated)
Cluster 2 and Cluster 3: 6391.9333 (distinct)
Cluster 2 and Cluster 4: 2959.0285 (moderately close)
Cluster 3 and Cluster 4: 3432.9048 (distinct but closer)

These clusters are distinct but closer compared to Cluster 1 and Cluster 2, suggesting some overlap.

These clustering results provide insights into the underlying data structure, informing future analyses and decision-making. Understanding the relationships among clusters enhances the interpretation of user behavior and resource utilization within the educational framework.

5 Data model use

The text part provides a thorough understanding of the machine learning model training process, the importance of hyperparameters, the challenges of overfitting, and the steps taken to improve model performance. It emphasizes the need for data quality and careful feature selection to achieve better outcomes in machine learning tasks.

Machine learning is the process that an algorithm goes through when trying to search for patterns in a massive amount of data. The ML advantage is that the algorithm tries to find the patterns automatically, it learns through the experience it gets from the data that gets fed into it. Mathematically, the program tries to find the best possible function to map a certain input to a certain output [34].

ML relies on examples of past experiences in the form of data to offer predictions with varying levels of confidence in the future. If an experience in the future very closely resembles an example seen by a ML model in the past, it can predict an outcome (also referred to as a target) with fair confidence. That space of ML with an object to predict is called supervised learning [35].

5.1 Data model training

Model training in ML is the process of teaching a machine learning algorithm to make predictions or decisions based on data.

Key components of model training:

Data is the foundation of machine learning, consisting of input features (attributes) and their corresponding output labels or targets. The model identifies patterns within this training data to make accurate predictions.

An algorithm serves as the mathematical or computational process that the model follows to learn from the data, establishing a set of rules to guide the learning process.

Parameters are internal settings or weights that the model fine-tunes during training based on the data. These adjustments help reduce the gap between the model's predictions and actual outcomes.

The loss function quantifies the error between the predicted values and actual target values. The model aims to minimize this error during training to improve its accuracy.

Optimization involves iteratively refining model parameters. Gradient descent, a widely used technique, helps identify the optimal values for these parameters. The compiled data model was trained using Python and PyTorch algorithms library in Colab Notebook [36].

Training a machine learning model involves several parameters, often categorized as hyperparameters (Table 2) and model parameters.

Table 2 Model hyperparameters

Hyperparameters	Aggregation
Epochs	10/100
Batch size	32
Hidden layers	5
Loss function	BCELoss
Optimizer	Adam
Learning rate	0.001
Activation functions	ReLU (Rectified Linear Unit)

Model parameters are the variables in a model that are learned from the training data. For instance, in neural networks, these include weights and biases. Typically, model parameters are not explicitly set before training but are initialized and then adjusted during the training process. For simple models: input size = 10, hidden size = 5, output size = 1 (for binary classification).

Hyperparameters are set before training and are used to control the learning process (Fig. 12).

Model training results (Fig. 12) explaning:

Training Loss: This decreases over time, starting at 0.7117 at epoch 10 and dropping to 0.6727 by epoch 100. This means the model is improving its fit to the training data, reducing errors as training progresses.

Validation Loss: However, the validation loss (performance on unseen data) increases from 0.6884 at epoch 10 to 0.6951 at epoch 100, indicating the model is performing worse on new data as training continues.

This shows that while the model is learning to perform better on the training data, it struggles to generalize to new, unseen data, suggesting overfitting. Overfitting occurs when the model memorizes the training data instead of learning general patterns.

Even when the number of training epochs was increased to 200, 500, and 1000, overfitting persisted. A potential solution is early stopping, where training is halted once the validation loss stops improving, preventing the model from learning noise in the data.

The model evaluation's data was checked with threshold 0.3 and 0.5 and the last one results are:

Unique values in predictions with threshold 0.5: tensor ([0, 1]).

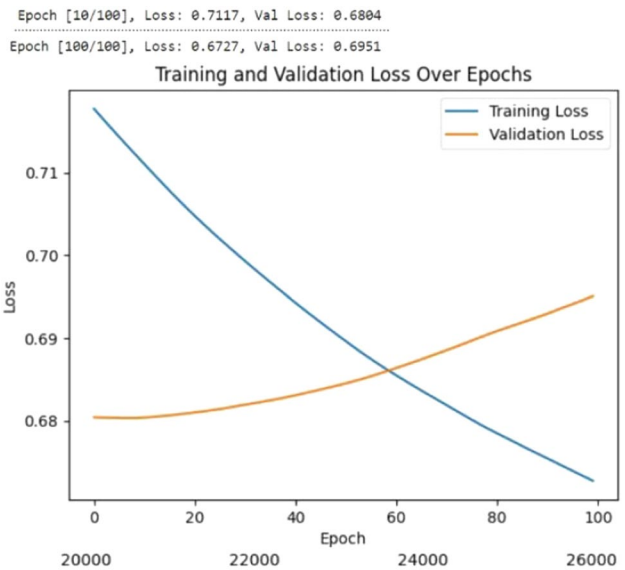


Fig. 12 Model training results and visualization

	Precision	Recall	F1-score	Support
Class 0	0.37	0.43	0.40	47
Class 1	0.41	0.36	0.38	53
Accuracy			0.39	100
Macro avg	0.39	0.39	0.39	100
Weighted avg	0.39	0.39	0.39	100

Class-wise Metrics

Class 0:

Precision (0.37): Out of all instances predicted as class 0, 37% were correctly classified.

Recall (0.43): Out of all actual instances of class 0, 43% were correctly identified by the model.

F1-score (0.40): The harmonic means of precision and recall, indicating a balance between them for class 0.

Class 1:

Precision (0.41): Out of all instances predicted as class 1, 41% were correctly classified.

Recall (0.36): Out of all actual instances of class 1, 36% were correctly identified by the model.

F1-score (0.38): The harmonic means of precision and recall, indicating a balance between them for class 1.

Overall Metrics

Accuracy (0.39): The overall accuracy of the model is 39%, meaning that 39 out of 100 instances were correctly classified.

Macro Average:

Precision (0.39): The average precision across all classes, treating each class equally.

Recall (0.39): The average recall across all classes, treating each class equally.

F1-score (0.39): The average F1-score across all classes, treating each class equally.

Weighted Average:

Precision (0.39): The average precision across all classes, weighted by the number of instances in each class.

Recall (0.39): The average recall across all classes, weighted by the number of instances in each class.

F1-score (0.39): The average F1-score across all classes, weighted by the number of instances in each class.

Results interpretation

Low Performance: The precision, recall, and F1-scores for both classes are low, indicating that the model is not performing well on this classification task.

Balanced Classes: Since the macro and weighted averages are the same (0.39), it suggests that the classes are relatively balanced in terms of the number of instances.

Improvement Needed: The overall accuracy of 39% is not much better than random guessing (50% for a binary classification problem), suggesting that the model may need significant improvements. Implementing strategies to improve data quality, feature engineering, model selection, and handling class imbalance can help enhance the model's performance.

To improve the performance of the 0.39 model, it is advisable to increase the data by aggregating, improve the quality of the input features by creating new features or selecting the most suitable ones.

5.2 Data model improving

Changing the hyperparameters and increasing the epochs from 200 to 1000 were used to improve the model training results, but no improvement was achieved, the model overfitted [36].

Since one of the possible reasons for the low efficiency of the model may be the influence of noise from extra data, the data were additionally cleared from results 0 in column log_count [37]. Gained results are:

Unique values in predictions with threshold 0.5: tensor ([0, 1]).

	Precision	Recall	F1-score	Support
Class 0	0.47	0.54	0.50	13
Class 1	0.60	0.53	0.56	17
Accuracy			0.53	30

	Precision	Recall	F1-score	Support
Macro avg	0.53	0.53	0.53	30
Weighted avg	0.54	0.53	0.54	30

The second set of results shows a noticeable improvement in model performance compared to the first set. Accuracy increased from 39 to 53%, and all classes and average metrics showed significant improvements. This indicates that the model accuracy, recall, and F1-score for both classes are significantly improved. However, it is important to note the smaller sample size in the second set, which may affect the stability of the results. Further testing using a further processed data set would be useful for a more complete assessment.

The log data sets were used to create a new data set in which the student ID was used to aggregate the new attributes using students ID code, Excel functions and teacher ratings.

ID (identification code) was introduced in Estonia on October 12, 1989. Each of the 11 digits of the identification code has a meaning associated with its owner [38].

Gender—the first digit of the personal identification number contains information about the owner's gender (odd number—male, even number—female) and for new data set: 1 is male and 0—female.

Age—the second and third digits of the year of birth were used for a student's age calculation.

Log_lesson—a student's logs number at lesson time.

Log_practice—a student's logs number at other time.

Logs—sum of log_lesson and log_practice of a student.

Mother_tongue—Estonian is 1 and Russian—0.

Grade—teacher ratings.

The new data set with dropped 'grade' column data was used in Python algorithms for these data prognosing.

6 Results analysis

The new data model was used in the interactive Colab Notebook for processing and training [39].

The use of the data from the grade column is needed to predict learning results. And it is important to have a forecast of the lack of learning outcomes to have preliminary information about failing students for timely intervention and prevention of student dropout from the learning process.

There is a high correlation between the data on the use of course resources for lectures and assessments—0.8, as well as the use of resources outside of lectures and assessments—0.68. Which is true—intensive use of educational resources affects the receipt of a high grade.

The low correlation of attributes of native language (0.17) and student gender (0.0047) with grade means that these parameters are not significant for learning success. The negative correlation of age attributes with a score of − 0.36 could mean that these parameters have an inverse relationship and with increasing age, the student's success decreases. In real life, this is not always the case, but in this group, this relationship is true.

The results table and the parameter correlation diagram are interesting and provide valuable information for the teacher regardless of his practical experience and can help to understand the dependencies of parameters and assessment for better planning of the educational process.

When training this model, the parameters were determined:

Binary Cross-Entropy Loss,

optimizer Adam,

lr = 0.001,

Epochs—10/100,

threshold = 0.01.

Training results are [38]:

Raw outputs (first 10): tensor ([0.0243, 0.0058, 0.0085, 0.0026]).

Unique values in predictions with threshold 0.01: tensor ([0, 1]).

	Precision	Recall	F1-score	Support
Class 0	1.00	0.75	0.86	4
Class 1	0.00	0.00	0.00	0
Accuracy			0.75	4
Macro avg	0.50	0.38	0.43	4
Weighted avg	1.00	0.75	0.86	4

Class 0.0:

Precision: 1.00—all instances predicted as class 0.0 were correct (no false positives).

Recall: 0.75—75% of actual class 0.0 instances were correctly identified by the model (one instance was missed).

F1-score: 0.86—the harmonic means of precision and recall, indicating a good balance between the two for class 0.0.

Support: 4—there were 4 actual instances of class 0.0 in the dataset.

Class 1.0:

Precision: 0.00—there were no predicted instances of class 1.0, so precision is 0.

Recall: 1.00—all actual instances of class 1.0 were identified correctly, but since there were 0 actual instances, this metric is not meaningful.

F1-score: 0.00—the F1-score is 0 because precision is undefined due to the lack of predicted instances for class 1.0.

Support: 0—there were no actual instances of class 1.0 in the dataset.

Accuracy: 0.75 (75%)—the model correctly classified 75% of the total instances.

Precision: 0.50—the average precision across both classes, treating each class equally.

The model has perfect precision for class 0.0, meaning no false positives, but the recall is 0.75, indicating that it missed 25% of the actual class 0.0 instances.

The overall accuracy of 75% is relatively high, but it is based on a very small sample size, which can lead to unreliable performance metrics.

The model shows a high accuracy of 75%, but this is based on a very small and imbalanced dataset. The precision for class 0.0 is perfect, but recall is moderate, indicating that the model missed some instances of class 0.0. The metrics for class 1.0 are not meaningful due to the lack of instances. Addressing the class imbalance and ensuring a larger, more balanced dataset for training and evaluation are crucial steps to improve the model's performance. More data is needed to ensure that both classes are adequately represented.

7 Conclusion and future work

The study emphasizes the importance of automating data collection and processing in education, using machine learning to predict student learning results, provide recommendations and develop personalized learning pathways.

Recording data during enrollment allows us to record information automatically and accurately without human intervention and avoid errors. Thus, automated collection and analysis of student data can help create a more flexible and adaptive education system that better matches the needs of each student/small group of students with specific characteristics to recommend and create e-learning pathways.

The study identified gaps in the preparation and processing of the obtained data—data handling requires automation to improve the accuracy of results and the speed of processing. The preliminary results of the study will be used to build and train new data models that can handle updated data and not just historical data over a certain period.

To obtain informative statistical data about educational resources on the Moodle platform and remove limitations, appropriate filters or the development and implementation of algorithms to collect the necessary information are required.

Using a course registration form with well-written questions is also useful for obtaining data that is not possible to obtain from available sources. Collecting enrolment data from other sources requires a process and is not easy, as it requires administrative permission to use the data for scientific purposes. However, this data can be useful for compiling personal characteristics that influence learning trajectories, such as gender, age, and native language. Personal data requires coding in accordance with the law on the protection of personal data. Identifying the type of data that can significantly influence learning trajectories is important in the process of collecting and preparing data.

In this study, the use of deep learning algorithms in processing existing enrolment data and learning resources for e-courses on the Moodle learning platform demonstrated relationships between the data that can be used for automated data processing and training a data model revealed ways to improve data preparation that can be used to predict learning success/failure.

Future research will focus on automating data preparation to improve the accuracy and refinement of models of existing enrolment data, as well as to ensure rapid provision of recommendations on more appropriate learning paths for students based on changes in their knowledge and learning outcomes.

In the next stage of the research work on the development of an automated personalized learning path, the author plans to continue research on the effectiveness of using student attributes using machine learning methods, develop an electronic course registration form with automated collection and processing of student registration data and use of the information obtained to provide personalized recommendations for the student. When determining student attributes, it is necessary to take into account factors that affect student engagement. Factors that contribute to development include positive student emotions, positive learning behavior, positive teacher behavior, teacher-student relationships and partnerships, students' ability to learn and think, supporting learning resources, individual and personal characteristics of students, learning factors [40]. Therefore, it is necessary to develop and use channels for collecting student feedback and processing it when using educational resources on the Moodle platform.

- TTK UAS processing of personal data—https://www.tktk.ee/isikuandmete-tootlemine/
- TTK UAS Moodle Terms of Use—https://moodle.tktk.ee/admin/tool/policy/viewall.php?returnurl=https%3A%2F%2Fmoodle.tktk.ee%2Fcourse%2Fview.php%3Fid%3D1295

Acknowledgements The author is grateful to Rytis Maskeliūnas, Professor at Kaunas University of Technology, for a detailed explanation of deep learning algorithms for use in scientific research and to Simona Ramanauskaitė, Professor at Vilnius Gediminas University of Technology, Lithuania, for her comments that contributed to the revision of the article.

Authors contributions O.O. wrote the main manuscript text and prepared all figures and tables, reviewed the manuscript.

Funding Not applicable.

Data availability All materials used for the article are openly available for use. The author provides open access for the use of processed data uploaded to the platform GitHub and available via the link.

Code availability The author provides open access for using the code in an interactive Colab Notebook, uploaded to the GitHub platform and available at the References link.

Declarations

Ethics approval and consent to participate The study was conducted in accordance with the guidelines of the Ethics Committee specified in the Ethics Statement and in accordance with the guidelines and regulations of the TTK UAS, Tallinn, Estonia:

Informed consent During the research, the rules for the protection and use of personal data were strictly observed. The author confirms that during the process of preparing the data for building and training the model, all personal information was replaced by numbering, which eliminates the possibility of identifying the individual.

Competing interests The authors declare no competing interests.

References

1. Howard SK, Swist T, Gasevic D, Bartimote K, Knight S, Gulson K, Apps T, Peloche J, Hutchinson N, Selwyn N. Educational data journeys: Where are we going, what are we taking and making for AI? Comput Educ Artif Intell. 2022;3:100073. https://doi.org/10.1016/j.caeai.2022.100073.
2. Maphosa V, Maphosa M. Fifteen years of recommender systems research in higher education: current trends and future direction. Appl Artif Intell. 2023;37(1): e2175106. https://doi.org/10.1080/08839514.2023.2175106.
3. Gomede E, de Barros RM, Mendes LDS. Deep auto encoders to adaptive E-learning recommender system. Comput Educ Artif Intell. 2021;2:100009. https://doi.org/10.1016/j.caeai.2021.100009.
4. General Data Protection Regulation. 2018. https://gdpr-info.eu. Accessed 15 May 2024.
5. Hussain M, Zhu W, Zhang W, Abidi SMR. Student engagement predictions in an e-learning system and their impact on student course assessment scores. Comput Educ. 2018;121:61–77.
6. Romero C, Ventura S. Data mining in education. Wiley Interdiscip Rev Data Min Knowl Discov. 2020;10(3): e1344.
7. Worsley M, Blikstein P. Leveraging multimodal learning analytics to understand cognitive processes in complex learning environments. J Learn Anal. 2018;5(2):30–50.
8. Tsai YS, Moreno-Marcos PM, Muñoz-Merino PJ. Learning analytics in higher education: a review of tools and practices. J Educ Technol Soc. 2018;21(4):16–27.
9. Lamas M, Vitorino R, Lamas A. Optimization in e-learning systems through artificial intelligence techniques. J Artif Intell Educ. 2019;30(2):183–204.
10. Wang Z, Yu L, Su B. The impact of personalized learning paths on students' academic performance in online education. Int J Educ Technol High Educ. 2020;17(1):1–12.
11. Imamah ULY, Arif D, Mauridhi HP. Enhancing students performance through dynamic personalized learning path using ant colony and item response theory (ACOIRT). Comput Educ Artif Intell. 2024;7:100280. https://doi.org/10.1016/j.caeai.2024.100280.
12. Hasegawa K, Tsukahara T, Nomiyama T. Associations between long-term care-service use and service- or care-need level progression: a nationwide cohort study using the Japanese Long-Term Care Insurance Claims database. BMC Health Serv Res. 2023;23:577. https://doi.org/10.1186/s12913-023-09615-0.
13. Fernandes AAA, Koehler M, Konstantinou N, Pankin P, Paton NW, Sakellariou R. Data preparation: a technological perspective and review. SN Comput Sci. 2023;4:425. https://doi.org/10.1007/s42979-023-01828-8.
14. Sharif A. What is data logging?. 2022. https://www.crowdstrike.com/cybersecurity-101/observability/data-logging. Accessed 20 Mai 2024.
15. Keskin S, Aydın F, Yurdugül H. The determining of outliers on e-learning data in the context of educational data mining and learning analytics. Educ Technol Theory Pract. 2019. https://doi.org/10.17943/etku.475149.
16. Landauer M, Onder S, Skopik F, Wurzenberger M. Deep learning for anomaly detection in log data: a survey. Mach Learn Appl. 2023;12:100470. https://doi.org/10.1016/j.mlwa.2023.100470.
17. Keskin S, Yurdugül H. E-learning experience: modeling students' e-learning interactions using log data. J Educ Technol Online Learn. 2022;5(1):1–13.
18. What is Data Logging? Blog. 2024. https://www.logmore.com/post/what-is-data-logging#:~:text=Data%20from%20the%20logging%20process,%2C%20and%20sensor(s). Accessed 23 May 2024.
19. GeeksforGeeks, computer science resources platform. Data Preprocessing in PyTorch. 2008. https://www.geeksforgeeks.org/data-preprocessing-in-pytorch. Accessed 24 May 2024.
20. Moodle architecture. 2018. https://docs.moodle.org/dev/Moodle_architecture. Accessed 22 May 2024.
21. Ganguly A, Student ME. A brief survey on issues & approaches of data cleaning. 2016. https://api.semanticscholar.org/CorpusID:212558418. Accessed 24 May 2024.
22. Ovtsarenko O. Logs data. San Francisco: GitHub; 2024.
23. Haque S, Mengersen K, Barr I, Wang L, Yang W, Vardoulakis S, Bambrick H, Hu W. Towards development of functional climate-driven early warning systems for climate-sensitive infectious diseases: statistical models and recommendations. Environ Res. 2024;249:118568. https://doi.org/10.1016/j.envres.2024.118568.
24. Roberts W. Blog post. Understanding the difference between deep learning and machine learning. 2024. https://ushur.com/blog/understanding-the-difference-between-deep-learning-and-machine-learning/. Accessed 26 May 2024.
25. IBM. Point charts, chart types, Cognos analytics. 2024. https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=types-point-charts. Accessed 17 July 2024.
26. Ovtsarenko O. Logs data correlation. San Francisco: GitHub; 2024.
27. Teachers Institute. Understanding different types of correlation in educational research. https://teachers.institute/assessment-for-learning/types-correlation-educational-research/. Accessed 18 July 2024.
28. GeeksforGeeks. Computer science resources platform. Types of Diagrams. 2008. https://www.geeksforgeeks.org/types-of-diagrams/. Accessed 24 May 2024.
29. Zaveri A. Unlocking the power of cluster analysis. 2023. https://mindthegraph.com/blog/cluster-analysis/. Accessed 30 May 2024.
30. Ovtsarenko O. Log data hierarchical clustering. San Francisco: GitHub; 2024.
31. Jaiswal S. K-means clustering algorithm. Free learning platform for better learning. 2021. https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning. Accessed 17 July 2024.
32. Ovtsarenko O. Log data K-means clustering. San Francisco: GitHub; 2024.
33. Abbey R. How to evaluate different clustering results. Paper 3409-2019. SAS Institute Inc.; 2019. https://support.sas.com/resources/papers/proceedings19/3409-2019.pdf. Accessed 22 July 2024.
34. Jain S. GeeksforGeeks post. 2024. https://www.geeksforgeeks.org/getting-started-with-classification/. Accessed 18 May 2024.
35. Nagpal M. How to train a machine learning model: the complete guide. ProjectPro Blog; 2024. https://www.projectpro.io/article/training-a-machine-learning-model/936. Accessed 20 May 2024.

36. Ovtsarenko O. Logs data model training. San Francisco: GitHub; 2024.
37. Ovtsarenko O. Updated data model training. San Francisco: GitHub; 2024.
38. Republic of Estonia E-resident. How to use your digital ID. 2024-09-03 07:26:47 UTC; 2024. https://learn.e-resident.gov.ee/hc/en-us/articles/360000624498-How-to-use-your-digital-ID. Accessed 13 July 2024.
39. Ovtsarenko O. New data model training. San Francisco: GitHub; 2024.
40. Li J, Xue E. Dynamic interaction between student learning behaviour and learning environment: meta-analysis of student engagement and its influencing factors. Behav Sci. 2022;13(1):59. https://doi.org/10.3390/bs13010059.