

An Artifact-Centric Process Mining Approach for Learning Style Analysis

Jiehua LU^{1,*}, Jun LI²

¹Dept of Internet of Things Technology, Hangzhou Polytechnic, Hangzhou, Zhejiang, China

²Information Engineering College, Hangzhou Dianzi University, Hangzhou, Zhejiang, China

Abstract

As an integrated discipline encompassing data mining, machine learning, process modeling and analytics, process mining is increasingly being applied in the field of education and has emerged as a prominent research topic. Traditional business process modeling approaches, which are primarily based on control flow rather than data flow, exhibit a limited capacity to capture a holistic view of critical business data within complex business procedures. This study focuses on the impact of data-driven process modeling techniques on the performance of analytical models and proposes an artifact-centric process mining approach for learning style analysis. Based on the artifact life-cycle model, we extracted sequences of data attribute operations that encapsulate learning style features. The similarity among different data attribute operation sequences was quantified. The proposed method was evaluated using the OULAD, a benchmark dataset in the learning analytics domain. Experimental results demonstrate that the method effectively enhances the performance of learning style prediction models, with SVM and GBoost algorithms outperforming other modeling approaches.

Keywords: Artifact, Process Mining, learning style analysis, Data Attribute Operation sequence, OULAD.

Received on 25 September 2025, accepted on 19 November 2025, published on 26 November 2025

Copyright © 2025 Jiehua LU *et al.*, licensed to EAI. This is an open access article distributed under the terms of the [CC BY-NC-SA 4.0](#), which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/eetsis.10390

1. Introduction

In recent years, the rapid development of digital education has enabled online learning platforms such as MOOCs and LMS to accumulate a vast amount of accessible educational data. This availability of data offers new opportunities to analyze students' behavioral patterns and learning processes. As a result, learning analytics (LA) has emerged as a significant field within educational research.

LA encompasses “the measurement, collection, analysis, and reporting of data related to learners and their contexts”, with the goal of enhancing the understanding of learning processes and improving the environments in which they occur[1]. It is closely connected to academic analytics, action analytics, and predictive analytics. Since the concept of LA was formally introduced in 2011, it has developed

into an interdisciplinary domain that combines educational theory, data science, and artificial intelligence [2]. Over recent years, the use of learning analytics in online educational settings has attracted growing interest from researchers, with the analysis of learning styles becoming one of the prominent areas of investigation [3].

The digital transformation of learning environments is driving a shift in learning style identification from static questionnaires toward dynamic, data-driven methodologies. Accordingly, the data-driven identification and application of learning styles have become a central focus and a growing trend in learning analytics research. Traditional learning styles theory, due to its reliance on static questionnaires and self-reported data, suffers from inherent limitations in capturing dynamic and contextualized learning processes. These limitations hinder

*Corresponding author. Email: ljh@mail.hzpt.edu.cn

its adaptation to digital learning environments and pose significant challenges to its predictive validity. The evolution in business process modeling from a "control-centric" to a "data-first" paradigm provides a new methodological pathway to address this dilemma. This shift moves the field from static inferences based on discrete surveys and subjective judgments toward a process-oriented characterization based on continuous behavioral logs and objective data, thereby establishing a viable path to fundamentally enhance the precision and timeliness of learning style analysis.

Building upon this premise, this study proposes an artifact-centric process mining framework for learning style analysis that coherently integrates data and process perspectives. Guided by the artifact life-cycle model, we extracted sequences of data attribute operations that embed learning style features. The similarity between these sequences was then measured to facilitate subsequent analysis. We conducted experiments using the Open University Learning Analytics Dataset (OULAD). Several algorithms, including GBoost, Random Forest, Logistic Regression, Support Vector Machine, and K-Nearest Neighbors, were employed to construct predictive models. Results indicate that the proposed approach significantly enhances the accuracy of learning style prediction.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the methodology of the proposed framework. Section 4 details the artifact-centric learning process modeling. Section 5 presents the learning style analysis algorithms based on the artifact life-cycle. Section 6 provides the experimental results and discussion. Finally, Section 7 concludes the paper and outlines future work.

2. Related work

This section opens with a survey of recent studies addressing various methods for analyzing students' learning styles in online learning environments. Subsequently, it examines current research trends within the domain of educational process mining.

2.1. Learning style analysis

With the development and growing maturity of online learning environments, which facilitate the large-scale collection of behavioral data, learning style analysis has garnered significant scholarly interest. The concept of learning style was first proposed by Herbert Thelen in 1954, referring to the distinctive personal characteristics that learners display when interacting with learning tasks [3]. Since its introduction, numerous related theories and models have emerged.

Classical learning style models—such as Kolb, FSLSM, and VARK—have been extensively applied and continue to evolve within online learning environments. Kolb's Learning Styles theory classifies learners into four distinct

categories: converging, diverging, assimilating, and accommodating [4]. FSLSM evaluates learning styles through four dimensions: information processing, perception, input, and understanding [5]. Among the most widely adopted frameworks in online education is the VARK model, which categorizes learning preferences into four sensory modalities: visual, auditory, reading/writing, and kinesthetic [6].

The digital transformation of learning environments is shifting learning style identification from static questionnaires to dynamic, data-driven approaches. Traditional theory's dependence on static data is fundamentally inadequate for modeling dynamic and contextualized learning, severely limiting its predictive power and applicability in contemporary educational settings. Consequently, data-driven identification and application of learning styles have become a major focus and emerging trend in learning analytics. On the one hand, researchers are increasingly turning to analysis techniques such as data mining, neural networks, and heuristic rule-based methods to analyze learning behaviors captured in online educational settings [7–9]. A significant challenge in this area, however, stems from the prevalence of unstructured data—such as text, images, audio, and video—which tend to be heterogeneous and incomplete. This complicates subsequent processes of data fusion, analysis, and interpretation [10]. On the other hand, advancements in business process modeling, which serves as a foundational framework for building learning style analysis systems, pave the way for capturing and analyzing behavioral feature data through process mining techniques.

2.2. Educational process mining

Serving as a bridge between data science and process science [11], Process Mining (PM) has been widely adopted across various sectors such as manufacturing and finance to improve overall performance. In the educational domain, the application of PM, referred to as Educational Process Mining (EPM), constitutes a specialized branch [12–13].

Educational Process Mining (EPM) applies data mining to historical educational data to analyze behavior and predict performance, while simultaneously constructing process models from event logs to generate insights for education management [14]. The primary technique involves discovering clear process models from event logs recorded by online learning platforms like MOOCs and LMS. The resulting models can accurately reconstruct learners' knowledge acquisition paths, map their resource interactions, and clarify their strategic choices, thereby revealing implicit behavioral patterns and learning characteristics. EPM provides an innovative approach to learning style analysis. By employing process mining on educational datasets, several studies have identified learning behavior patterns and developed predictive models based on historical learning behaviors [15–17].

Within EPM, the process occupies a central role, process modeling is essential to overall performance [18]. Traditional process modeling approaches are primarily based on control flow rather than data flow, which limits their ability to capture a comprehensive view of critical business data within complex procedures. Currently, business process modeling has shifted from a process-centric(control-centric) view to a data-centric(data-first) perspective. A representative example of this evolution is the emergence of artifact-centric approaches to business process modeling [19-20]. An artifact is defined as a specific, identifiable, and self-descriptive unit of information within a process that encapsulates all data necessary for process completion [21]. Artifact-centric business process models emphasize business-relevant data objects and their lifecycles, placing particular focus on the data aspects of processes [22]. Currently, some researchers analyze learning-generated artifacts to evaluate both learner behaviors and external interventions, thereby informing their understanding of learning processes[23-25].

In summary, the adoption of artifact-centric process mining enhances the efficiency of extracting meaningful learning information, leading to improved understanding of learning activities, more effective analysis of learning styles, and more accurate recommendation generation. Within this approach, artifacts encapsulate relevant data objects and their lifecycles, recording which tasks trigger state transitions in the data. The entire business process model—including its design, analysis, and system implementation—revolves around these artifacts. This methodology offers a novel solution to the aforementioned challenges in learning analytics: unstructured learning style data can be standardized and encapsulated within corresponding artifacts, while personalized learning styles can be tracked by monitoring the lifecycle of each individual artifact.

3. Methodology

This section delineates the methodology for developing an artifact-centric process model for online learning(ACPM-OL), which serves as the core analytical framework for our learning style analysis. The architecture, as illustrated in fig. 1, comprises three sequential stages: (1) Learning Feature Extraction and Artifact Definition, (2) Artifact-Centric Model Construction, and (3) Learning Style Analysis.

3.1 Learning feature extraction and artifact definition

The process initiates with data preprocessing and feature engineering using the OULAD dataset. A set of pivotal learning features was selected to define the data attributes of our core artifacts. For instance, features related to learning interactions—such as `sum_click` (total interactions), `activity_type`, and temporal markers

(`day_from`, `day_to`)—were mapped directly to define the Interactive Artifact (see Table 1 for a complete listing). Similar procedures were applied to define the StudentInfo Artifact and Assessment Artifact, ensuring all data attributes are grounded in the observable learning behaviors from the dataset.

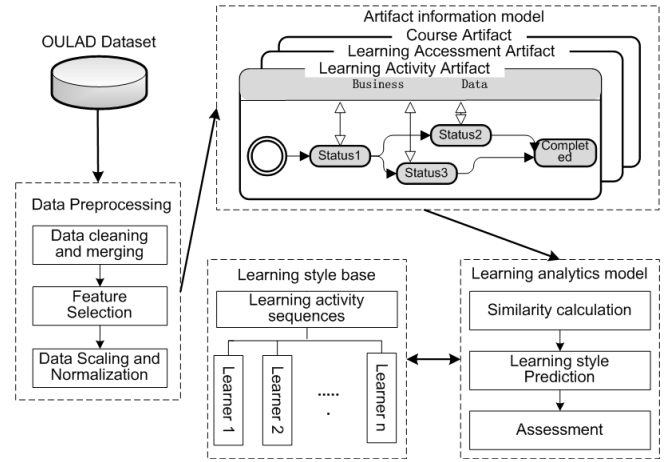


Fig.1. Artifact-centric process mining approach for learning style analysis model

3.2. Artifact-centric model construction

Subsequently, the ACPM-OL was formally constructed as a tuple (A, L, R) , where A is a set of artifact classes, L is a set of learning activities, and R is a set of business rules (see Section 4 for formal definitions). For each artifact class, a finite-state machine was designed to represent its lifecycle, capturing the permissible sequence of states (S) triggered by learning activities. Business rules (λ , β , L) were then formulated to explicitly link these activities to state transitions. This structured approach transforms raw event data into a formal, executable process model.

3.3. Learning style analysis based on artifact lifecycles

The final stage leverages the constructed ACPM-OL for learning style analysis. We propose a novel similarity metric that operates on artifact lifecycle instances. Specifically, for each student, we extract the sequence of data operations (O) from their artifact lifecycles (as per Definition 6). The similarity between two students is computed based on the normalized compression distance (NCD) between their respective data operation sequences. This method effectively captures behavioral patterns encoded in the artifact state transitions. The resulting similarity matrix is then clustered using the K-means algorithm and others to identify distinct learning style groups, which are subsequently cataloged in a learning-style dictionary.

4. Artifact-centric learning process modeling

This section presents a method for artifact-centric learning process modeling and introduces an artifact-oriented view for data-centric online learning processes. The relevant formal definitions are established herein.

4.1. An artifact view for business processes

In the artifact-centric learning process model, an artifact refers to a data entity that participates in the learning process. Each artifact comprises a set of data attributes and distinct states. Transitions between states are triggered by predefined business rules. The progression of an artifact from its initial state to its final state constitutes its lifecycle. Throughout this lifecycle, data attributes are continuously read, updated, inserted and deleted in response to learning behaviors.

Definition 1: Artifact Class. An artifact class represents a category of artifacts that share common data attributes and state configurations. Formally, an artifact class A is defined as a tuple (A, S) , where:

- $A = \{a_1, a_2, \dots, a_n\}$, $a_i (1 \leq i \leq n)$ is an data attribute;
- $S = \{s_1, s_2, \dots, s_m\}$, $s_i (1 \leq i \leq m)$ is a state.

For example, Table 1 presents the data attributes of the interactive artifact within the learning process.

Table 1. Data attributes and descriptions of interactive artifact

Attribute	Description
ID_Interactive	A unique identifier for the instance of interactive learning behavior
ID_Stu	A unique student identifier
ID_Site	A unique identifier for the VLE material
Activity_Type	The role type assigned to the material during student interaction
Day_from	The start date of the material's current use
Day_to	The end date of the material's current use
Sum_Click	The total count of student interactions of the material's current use

Definition 2: Artifact instance. Given artifact class A , we use a four tuple (id, A, s, a) to express one operate instance of this class, id is the unique identifier for this instance, s is current state, $a = \{a_1, a_2, \dots, a_n\}$ denote each data attribute's value of current state.

Definition 3: Learning activity. A learning activity refers to a data operation within an online learning context that consumes input from one or more artifacts and generates an output through the modification of artifact states.

Definition 4: Business rules. Business rules provide explicit representations of the internal business logic within a learning process. Business rule R can be defined as a triple (λ, β, L) where, λ and β denote the pre-condition and post-condition respectively—representing the input and output data of the rule—that collectively trigger its execution, and L corresponds to a learning activity to be carried out.

Each business rule encapsulates a segment of the business processes in online learning, specifying which learning behavior is activated and which artifact states are altered, based on specified pre- and post-conditions. A single learning behavior may trigger multiple business rules and involve several artifacts.

Table 2 presents a business rule refers to changes of states of multiple artifacts. The rule $R1$ changes for the Studentinfo artifact from the course selected state to the VLE interacted state, and simultaneously changes from the started state to the finished state for the Interactive artifact.

Table 2. Example of a business rule involving multiple artifacts

rule name	R1
related artifacts	Studentinfo, Interactive
λ (pre-condition)	Instate(Studentinfo, course selected) Read(Studentinfo.current_course) Read(Studentinfo.ID_stu) Instate(Interactive, started) Read(Interactive.ID_site)
β (post-condition)	Instate(Studentinfo, VLE interacted) Insert(Studentinfo.ID_Site) Instate(Interactive, finished) update(Interactive.ID_stu) update(Interactive.activity_type) update(Interactive.day_from) update(Interactive.day_to) update(Interactive.sum_click)
L (learning activity)	An interaction with a learning resource

Definition 5: Artifact-Centric Process Model for Online Learning, ACPM-OL. We use a triple-tuple (Z, L, R) to express it, where:

— $Z = \{Z_1, Z_2, \dots, Z_n\}$, $Z_i (1 \leq i \leq n)$ is a Artifact instance associated with this process;

— $L = \{L_1, L_2, \dots, L_m\}$, $L_i (1 \leq i \leq m)$ is a learning activity for reading or writing operations in Artifact attributes data;

— $R = \{R_1, R_2, \dots, R_k\}$, $R_i (1 \leq i \leq k)$ is a business rule which trigger data interaction and Artifact state transition.

fig. 2 illustrates the interactive learning process from an artifact-oriented perspective within an online learning environment, as part of the ACPM-OL. The process involves three artifacts: the StudentInfo Artifact, the Interactive Artifact, and the Assessment Artifact. Each rounded rectangle represents a state of an artifact instance, and each arrow indicates a business rule that corresponds to a state transition. Hollow arrows denote business rules that involve multiple artifacts. As shown, rule R1 triggers simultaneous state transitions in both the StudentInfo Artifact and the Interactive Artifact. Concurrently, corresponding data operations are activated—updating student information data stored in the StudentInfo Artifact and recording VLE interaction data within the Interactive Artifact.

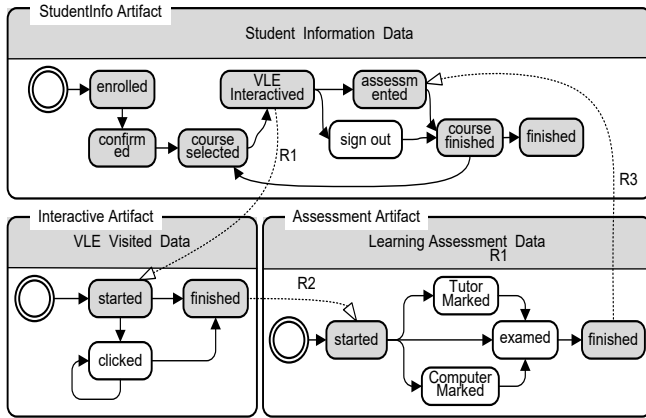


Fig.2. An Artifact view of learning interactive process

4.2. Artifact life-cycle

Each artifact possesses a lifecycle that depicts the evolution of its states throughout the learning process. During the Artifact life-cycle, data attributes will perform corresponding data operations on each state change. Meanwhile, due to the complex control structure of the business process, the same data attribute may be recorded synchronously in multiple artifacts.

Definition 6: Artifact Life-cycle Schema. An Artifact Life-cycle Schema provides a formal representation of the state transitions inherent to an artifact class, we use a triple- (Z, T, O) to express it, where:

— Z is a artifact instance of artifact A , which holds its current state and values of data attributes.

— T represents an ordered set comprising all activities within a given artifact instance and serves as a ternary transition relation. $T \subseteq A.S \times A.R \times A.S$. A transition $t = (s_i, r_j, s_k) \in T$ indicates that the artifact instance transitions from a source state s_i to a target state s_k when the precondition of business rule r_j is satisfied.

— O represents the data operation performed by activity, An operation $o = (dm, tn, \rho)$ represents the data operation ρ

performed by activity tn on the artifact's data attribute dm , where $\rho \subseteq \{\text{read, write, add, remove}\}$.

The operations on data attributes triggered by learning activities can form a ordered data operation sequence, which also contains the behavioral characteristics of learning processes. Different learning activities create different artifact instances and generating different data operation sequences. Therefore, by extracting the data operation sequence of each artifact in ACPM-OL, the learning process can be decomposed into a set of sequences of data attribute operations.

fig. 3 shows the life-cycle schema of the interactive artifact. In the figure: t_0, t_1 represent the names of activities; below the activity name are the corresponding data operations. The bottom of the figure is the data slots for information gathered as an interactive artifact instance evolves, including the change of status attributes and details of learning behavior data attributes. During the evolution of learning activities, corresponding data operations are recorded in these data slots. Consequently, given the action trajectory of an artifact instance, the corresponding data operation sequence can be extracted from the Life-cycle Schema.

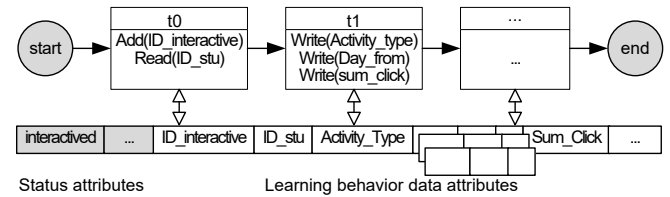


Fig.3. Life-cycle Schema of ACPM-OL: the interactive artifact

5. learning style analysis based on artifact life-cycle

In this section, learning styles are analyzed and predicted based on the artifact life-cycle model by comparing similarities among data attribute operation sequences that contain learning style features.

5.1. Similarity calculations

As previously described, each artifact possesses a lifecycle that depicts the sequence of activities and associated state changes throughout the learning process. There are certain dependencies between the states of different artifacts, the state transitions may be triggered simultaneously by a certain business rule. By using the artifact-centric choreograph approach[26], the states of different artifacts are captured and combined to form a complex state transition graph.

Based on the life-cycle schema of ACPM-OL, as learning activities evolve, the corresponding state transitions and data operations are recorded. Hence each

activity sequence has an equivalent state transition sequence and the corresponding data operation sequence .

Suppose that β_1 and β_2 are two activity sequences, use formula 1 to calculate their status similarity.

$$\text{sim_stat}(\beta_1, \beta_2) = \frac{\text{equal}(|\beta_1|, |\beta_2|)}{\max(|\beta_1|, |\beta_2|)} \quad (1)$$

Where $\text{equal}(|\beta_1|, |\beta_2|)$ denotes the number of matching states at corresponding positions in the two state sequences, $\max(|\beta_1|, |\beta_2|)$ refers the maximum number of states contained in the two state sequences.

use formula 2 to calculate their transfer similarity.

$$\text{sim_tran}(\beta_1, \beta_2) = \frac{\text{common}(|\beta_1|, |\beta_2|)}{\text{tran}(|\beta_1|) + \text{tran}(|\beta_2|)} \quad (2)$$

Where $\text{common}(|\beta_1|, |\beta_2|)$ refers to the frequency of identical state transition events occurring in both sequences, $\text{tran}(|\beta_1|) + \text{tran}(|\beta_2|)$ represents the total number of state transitions in β_1 and β_2 .

use formula 3 to calculate their data attributes similarity.

$$\text{sim_att}(\beta_1, \beta_2) = \frac{\sum_{i=1}^K \text{adapt}(|\beta_1|, |\beta_2|)}{K} \quad (3)$$

Where K is the number of data operations that read the same data attributes and performed the same data operation type, $\text{adapt}(|\beta_1|, |\beta_2|)$ is the semantic similarity of attribute assignment statements.

The similarity of activity sequences between β_1 and β_2 is derived by the combination of status similarity, transfer similarity and data attributes similarity, and its calculation formula is as follows:

$$\text{sim}(\beta_1, \beta_2) = \omega_1 \times \text{sim_stat}(\beta_1, \beta_2) + \omega_2 \times \text{sim_tran}(\beta_1, \beta_2) + \omega_3 \times \text{sim_att}(\beta_1, \beta_2) \quad (4)$$

where ω_1 , ω_2 , and ω_3 are weighting coefficients that sum to 1, with each being greater than or equal to zero.

5.2. Learning style assessment

Online learning generates a large amount of learning behavior data. In ACPM-OL, each learning behavior of every learner will produce a activity sequence, which constitutes their learning activity sequence set. If there are many activity sequences with high similarity in the activity sequence set of two learners, then these two learners have similar learning interest preferences and can be defined as having the same learning style.

Learning behaviors vary not only across different students but also within the same student over time. The set of learning activity sequences is often highly complex, with multiple potential paths from the initial to the final activity. Therefore, the evaluation of learning styles must incorporate temporal factors.

To assess the learning styles of two students (denoted as S_1 and S_2), the process consists of the following steps:

1. Extract all learning activity sequences for S_1 and S_2 ;
2. Calculate the frequency of each learning activity sequence. The most frequently occurring sequence is designated as the key activity sequence, denoted as β_1, β_2 , while the remaining sequences are classified as ordinary activity sequences, denoted as β_1', β_2' ;

3. Partition the ordinary activity sequences into N comparison groups according to time intervals, labeled as $(\beta_{1i}', \beta_{2i}')$, where $1 \leq i \leq N$;

4. Compute the similarity between the key activity sequence and each ordinary activity sequence group using Formula 4

5. Determine the similarity between the learning styles of the two students using Formula 5.

$$\text{sim}(S_1, S_2) = \omega_1 \times \text{sim}(\beta_1, \beta_2) + \omega_2 \times \frac{\sum_{i=1}^N \text{sim}(\beta_{1i}', \beta_{2i}') \times \text{fd}(t_i)}{\max(|\beta_1'|, |\beta_2'|)} \quad (5)$$

Where ω_1, ω_2 represent the weights of the key activity sequence and ordinary activity sequences respectively. Since the key activity sequence generally has a greater impact, so $\omega_1 > \omega_2$; $\text{fd}(t_i)$ is a time decay factor, where more recent activity sequences receive larger values; $\max(|\beta_1'|, |\beta_2'|)$ is the sum of all ordinary activity sequences.

6. Experiment and evaluations

In this section, we assessed the efficacy of the artifact-centric learning process modeling approach on the benchmark dataset OULAD.

6.1. Dataset

Datasets serve as foundational components in data science and machine learning, with representative and diverse data being essential for both educational data mining and learning analytics.

Widely regarded as one of the benchmark datasets in the field of learning analytics, OULAD's key feature is its combination of student demographic details with aggregated clickstream records from VLE interactions, which supports behavioral analysis of learning processes through fine-grained action data [27-28]. Originating from a curated subset of student records at the Open University UK (OUUK), OULAD encompasses seven distinct courses presented over several semesters. It incorporates demographic profiles and assessment scores for more than 30,000 students, along with detailed VLE interaction logs that compile daily click events—totaling over 10 million entries.

OULAD has been extensively used in educational data mining and learning analytics research, particularly in the development of predictive modeling frameworks [29]. The dataset offers a comprehensive range of educational variables, including demographic attributes, academic performance indicators, and behavioral traces such as click patterns, activity types, time engaged, and access frequency.

6.2 Preprocessing of data

OULAD consists of seven relational tables. But the dataset doesn't meet artifact's data collection requirements and

cannot be directly used by any machine learning algorithm. Data cleaning and integration were first performed to manage missing values and mitigate the impact of outliers. Second step, extract learning features by feature engineering and aggregate various features by id_student. The corresponding data attributes of artifacts are set on the basis of these features. Third step, according to the characteristics of artifacts' data attributes, the relational tables were aggregated and merged into a dataset that can be processed by ACPM-OL.

After merging the data, scaling and encoding procedures were applied. Scaling normalizes features to a $[0, 1]$ range to ensure comparability across different units and magnitudes, while encoding transforms categorical features into numerical representations compatible with machine learning models. Subsequently, applied the K-means algorithm to cluster the pre-processed dataset. By using a feature selection algorithm based on the improved elbow method and silhouette coefficient, the optimal K value was selected and the data was pre-grouped. After data preprocessing, seven relational tables are condensed and converted into three artifact tables.

6.3 Result and discussion

The data of OULAD is widely utilized to the research on learning analytics, especially focused on predictive modeling. The modeling algorithms predominantly employed include Logistic Regression, K-Nearest neighbour, random forest, Gradient-Boosted Tree[30-33].

Instead of using raw OULAD, The data preparation process produced a artifact-centric information model (ACPM-OL) which was more suitable for learning style analysis, especially for predictive models. The experiment is carried out led by the two research questions that were defined:

1. Does the ACPM-OL approach significantly enhance the predictive accuracy of learning style models compared to using the original dataset?

2. Among classical predictive modeling algorithms applied within the ACPM-OL framework, which one achieves the highest performance in predicting students' learning styles?

To address these questions, a series of experiments were conducted.

For the first research question, an experimental scenario was designed. The GBoost algorithm was employed as the classifier. Both the raw OULAD dataset and the ACPM-OL processed dataset were used as inputs. Two experimental trials were conducted using 80% to 90% of each student's data for training, with the remaining 10% to 20% reserved for testing. All three evaluation metrics were applied to assess model performance.

Table 3. GBoost Model performance under different input data

InfoModel	TestProportion	precision	recall	f1-score
Raw	10%	0.87	0.89	0.88
Raw	20%	0.85	0.83	0.86
ACPM-OL	10%	0.96	0.98	0.97
ACPM-OL	20%	0.96	0.93	0.94

Table 3 presents the experimental results obtained with different information models and varying test set sizes. The choice of information model is shown to significantly influence the predictive performance of the model. ACPM-OL model improves the data quality of the prediction model, significantly improving precision, recall, F1score. The stability of the prediction model is also enhanced, when the proportion of test set changes significantly, the performance index does not fluctuate greatly.

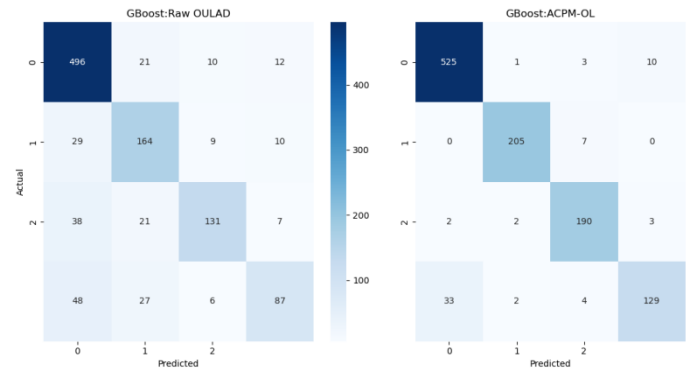


Fig.4. Confusion Matrix Heatmap of the GBoost model under different input data

Fig. 4 displays the confusion matrix heatmap. The left panel corresponds to prediction results using the raw OULAD dataset, while the right panel shows results obtained with the ACPM-OL processed data. This study considers four predefined learning styles. The diagonal entries indicate the number of correctly classified instances, and it is clearly evident that the model exhibits improved predictive accuracy when using the ACPM-OL framework.

Table 4. Model performance

Model	TrainAccuracy	TestAccuracy	TrainF1 score	TestF1 score
SVM	0.971429	0.934140	0.970735	0.933397
GBoost	1.000000	0.928763	1.000000	0.927910
Random Forest	0.908571	0.912634	0.906147	0.911465
Logistic Regression	1.000000	0.909946	1.000000	0.911040
KNN	0.830588	0.803763	0.823860	0.799008

For the second research question, a corresponding experimental scenario was constructed. The ACPM-OL dataset was used as input, the dataset was partitioned into a training set (90%) and a test set (10%) using a random stratified split. This approach ensures that the distribution of the target variable (i.e., the learning styles) is consistent across both subsets. The random seed was set to 42 for reproducibility. GBoost and other four classification algorithms were employed. As shown in Table 4, the SVM algorithm achieved the highest performance, with an accuracy of 93.4% and the highest F1-score. In contrast, the KNN algorithm significantly underperformed compared to the other models. Although the parameter `n_neighbors` was tuned, no substantial improvement was observed.

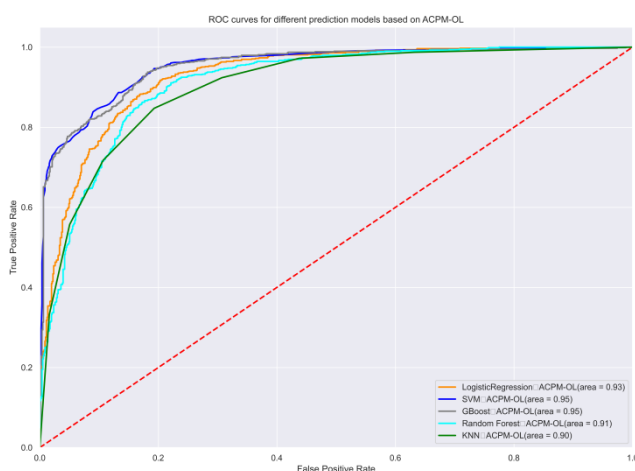


Fig.5. ROC curves for different prediction models based on ACPM-OL

The ROC curve serves as a widely used and intuitive metric for evaluating model performance. fig. 5 presents

the ROC curves of the five models mentioned above. All curves lie close to the top-left corner, indicating high true positive rates with low false positive rates, which suggests good classification capability. Among them, SVM and GBoost achieved the highest AUC values, both reaching 0.95, demonstrating superior overall discriminative ability.

7. Conclusion and future work

This study examines the impact of data-driven process modeling techniques on predictive performance by proposing an artifact-centric process mining approach. Based on the artifact life-cycle model, learning styles were analyzed and predicted by measuring similarity among sequences of data operations that encapsulate learning style features. OULAD was used for experimental validation. Results demonstrate that the proposed method significantly enhances the performance of learning style prediction models. Among the evaluated algorithms, SVM and GBoost exhibited superior predictive accuracy compared to other models. The multidimensional, continuous nature of the Felder-Silverman model aligns well with digital learning environments, while the VARK model offers unique utility for connecting with multimodal resources. Therefore, our future work will establish FSLSM as the core framework, integrating VARK to enhance our analytical dimensions and achieve more precise, dynamic learning style predictions.

Acknowledgements.

Thank the associate editor and anonymous reviewers for their constructive comments. This material is based upon work funded by Zhejiang Provincial Natural Science Foundation of China and Zhejiang Province's High-quality Online Open Course for continuing education Construction Project.

References

- [1] Mian YS, Khalid F, Qun AWC, Ismail SS. Learning analytics in education, advantages and issues: a systematic literature review. *Creative Education* 2022;13(9):2913-20.
- [2] Kew SN, Tasir Z. Learning analytics in online learning environment: A systematic review on the focuses and the types of student-related analytics data. *Technology, Knowledge and Learning* 2022;27(2):405-27.
- [3] Costa RD, Souza GF, Valentim RAM, Castro TB. The theory of learning styles applied to distance learning. *Cognitive Systems Research* 2020;64:134-45.
- [4] Saifi AG, Khlaif ZN, Affounch S. The effect of using community-based learning program in science students' achievement according to Kolb's learning styles. *Social Sciences & Humanities Open* 2024;10:101125.
- [5] Saringat MZB. A Systematic Literature Review Enhanced Felder Silverman Learning Style Models (FSLSM). In: *2022 Seventh International Conference on Informatics and Computing (ICIC)*; 2022 Dec. p 1-7.
- [6] Abomelha F, Newbury P. A VARK learning style-based Recommendation System for Adaptive E-learning. *Annals*

- of Computer Science and Information Systems 2024;41:1-8.
- [7] Lestari A, Lawi A, Thamrin SA, Hidayat N. Automated Detection of Learning Styles using Online Activities and Model Indicators. *International Journal of Advanced Computer Science & Applications*. 2024 Jun 1;15(6).
 - [8] Essa SG, Celik T, Human-Hendricks NE, et al. Personalized adaptive learning technologies based on machine learning techniques to identify learning styles: A systematic literature review. *IEEE Access* 2023;11:48392-409.
 - [9] Mahadevkar SV, Khemani B, Patil S, Kotecha K, Vora DR, Abraham A, et al. A review on machine learning styles in computer vision—techniques and future directions. *IEEE Access* 2022;10:107293-329.
 - [10] Shao M, Zhao L. Analysis of Learning Styles Based on Multimodal Technology: Evolution, Dilemmas and Scenarios. *Open Education Research* 2022;28(4):102-9.
 - [11] Van Der Aalst W. *Process Mining*. Berlin, Heidelberg: Springer; 2016.
 - [12] Bogarín A, Cerezo R, Romero C. A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2018;8(1):e1230.
 - [13] Semler RF, Semler JR, Wehrmeister MA, Southier LF, Casanova D, Teixeira M, et al. Educational process mining: literature classification, gaps, and emerging opportunities. *Frontiers in Education* 2025 Jun;10:1543761.
 - [14] Feng G, Chen H. Educational process mining: A study using a public educational data set from a machine learning repository. *Education and Information Technologies* 2025;30(6):8187-214.
 - [15] Feng G, Fan M, Ao C. Exploration and visualization of learning behavior patterns from the perspective of educational process mining. *IEEE Access* 2022;10:65271-83.
 - [16] Hachicha W, Ghorbel L, Champagnat R, Zayani CA, Amous I. Using process mining for learning resource recommendation: a moodle case study. *Procedia Computer Science* 2021;192:853-62.
 - [17] Ma F. Learning behavior analysis and personalized recommendation system of online education platform based on machine learning. *Computers and Education: Artificial Intelligence* 2025;8:100408.
 - [18] AlQaheri H, Panda M. An education process mining framework: Unveiling meaningful information for understanding students' learning behavior and improving teaching quality. *Information* 2022;13(1):29.
 - [19] Staudinger S, Schuetz CG, Schrefl M. Using Multilevel Business Artifacts for Knowledge Management in Analytics Projects. In: *2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*; 2023 Oct. p 689-98.
 - [20] Hillemacher S, Jäckel N, Kugler C, Orth P, Schmalzing D, Wachtmeister L. Artifact-based analysis for the development of collaborative embedded systems. *Model-Based Engineering of Collaborative Embedded Systems: Extensions of the SPES Methodology*. Cham: Springer International Publishing; 2020. p 315-31.
 - [21] Yongchareon S, Liu C, Zhao X. A framework for behavior-consistent specialization of artifact-centric business processes. In: *International Conference on Business Process Management*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012 Sep. p 285-301.
 - [22] Atouani A, Kirchhof JC, Kusmenko E, Rumpe B. Artifact and reference models for generative machine learning frameworks and build systems. In: *Proceedings of the 20th ACM SIGPLAN International Conference on Generative Programming: Concepts and Experiences*; 2021 Oct. p 55-68.
 - [23] Peppler K, Davis-Soylu HJ, Dahn M. Artifact-oriented learning: A theoretical review of the impact of the arts on learning. *Arts Education Policy Review* 2023;124(1):61-77.
 - [24] Trausan-Matu S, Slotta JD. Artifact analysis. In: *International handbook of computer-supported collaborative learning*. Cham: Springer International Publishing; 2021. p 551-67.
 - [25] Jahnke I, Meinke-Kroll M, Todd M, Nolte A. Exploring artifact-generated learning with digital technologies: Advancing active learning with co-design in higher education across disciplines. *Technology, Knowledge and Learning* 2022;27(1):335-64.
 - [26] Lohmann N, Wolf K. Artifact-centric choreographies. In: *International conference on service-oriented computing*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010 Dec. p 32-46.
 - [27] Alhakbani HA, Alnassar FM. Open learning analytics: a systematic review of benchmark studies using open university learning analytics dataset (oulad). In: *Proceedings of the 2022 7th International Conference on Machine Learning Technologies*; 2022 Mar. p 81-6.
 - [28] Kuzilek J, Hlosta M, Zdrahal Z. Open university learning analytics dataset. *Scientific Data* 2017;4:170171.
 - [29] Jin L, Wang Y, Song H, So HJ. Predictive modelling with the Open University Learning Analytics dataset (OULAD): A systematic literature review. In: *International Conference on Artificial Intelligence in Education*; 2024 Jul. p 477-84.
 - [30] Jin L, Wang Y, Song H, So HJ. Predictive modelling with the Open University Learning Analytics dataset (OULAD): A systematic literature review. In: *International Conference on Artificial Intelligence in Education*; 2024 Jul. p 477-84.
 - [31] Kukkar A, Mohana R, Sharma A, Nayyar A. A novel methodology using RNN+ LSTM+ ML for predicting student's academic performance. *Education and Information Technologies* 2024;29(11):14365-401.
 - [32] Lakshmi PR, Geetha AV, Priyanka D, Mala T. PRISM: Predicting student performance using integrated similarity modeling with graph convolutional networks. In: *2023 12th International Conference on Advanced Computing (ICoAC)*; 2023 Aug. p 1-7.
 - [33] Liu Y, Fan S, Xu S, Sajjanhar A, Yeom S, Wei Y. Predicting student performance using clickstream data and machine learning. *Education Sciences* 2022;13(1):17.