CIT. Journal of Computing and Information Technology, Vol. 34, No. 3, September 2026, 147–173
doi: 10.20532/cit.2026.1006450

Mahalanobis-enhanced K-Means++ Clustering for Multidimensional Learner Personas: Evidence from the Canvas MOOC Dataset

Yafeng Liu[1], Rui Yan[2], Yong Xiao[2], Ailing Chen[2] and Di Zhang[1]

[1] Further Education Department, Hengshui Open University, Hengshui, China
[2] University of Science and Technology Beijing, Beijing, China

Smart learning environments increasingly rely on data-driven methods to support individualized instruction. This paper proposes a multidimensional learner persona modeling framework that integrates behavioral, cognitive, and motivation-related indicators derived from LMS logs and self-report proxies. Using the Canvas open dataset, we construct five learner indicators (active days, number of interactions, course completion rate, user type, and learning expectations) and evaluate clustering performance under different distance metrics. We implement K-Means++ with Euclidean distance and a covariance-aware variant using Mahalanobis distance computed from the standardized feature covariance matrix, and compare them with density-based (DBSCAN) and probabilistic baselines. Results show that Mahalanobis K-Means++ yields improved clustering quality (Silhouette = 0.4277; Davies–Bouldin = 0.8118; Calinski–Harabasz = 1906.56) compared with Euclidean K-Means++ (Silhouette = 0.3281; Davies–Bouldin = 0.8279; Calinski–Harabasz = 1472.13). Four interpretable personas are identified and mapped to differentiated instructional interventions across task, activity, resource, and evaluation dimensions, with feasibility assessed via expert review. The study highlights the practical value of distance-metric selection for learner segmentation and provides a reproducible pipeline for persona-driven instructional support.

ACM CCS (2012) Classification: Computing methodologies → Machine learning → Unsupervised learning → Clustering

Keywords: learner persona; precision teaching; instructional intervention; K-Means++; smart teaching

1. Introduction

With the acceleration of the digital transformation of education, the smart teaching environment is reshaping the traditional teaching mode through the integration of big data, artificial intelligence and other technologies, bringing significant changes and innovations to education, and promoting the development of scientific decision-making, intelligent management, precise teaching, and personalized learning [1]. The introduction of MOOC resources, flipped classrooms, SPOC platforms, *etc.*, helps to make full use of high-quality online course resources and promotes change and innovation in the classroom teaching mode of colleges and universities. The core demand of this change is to break through the limitations of "standardized production" in traditional teaching and realize the paradigm shift from "teacher-centered" to "learner-centered" [2]. It has been shown that smart classrooms can significantly improve students' learning engagement and participation, and promote the healthy development of students' comprehensive ability. In addition, smart teaching can enhance students' motivation and interest in learning, improve teaching quality and efficiency, and inject new vitality into the development of higher education and the cultivation of innovative talents [3-5].

Precision teaching is a teaching method based on behavioral learning theory proposed by the

American scholar Ogden Lindsley in the 1960s. As the core practical path of intelligent education, it emphasizes the measurability of teaching goals and the precision of intervention strategies through high-frequency data collection and visual analysis, and has gradually developed into an effective framework for evaluating the teaching of any subject and any academic level [6]. Research has pointed out that precision teaching is an efficient teaching method that can effectively shorten learning time, promote learning outcomes, and provide students with personalized learning experiences [7].

As a data-driven core tool, learner personas can portray learners' behavioral, cognitive and emotional characteristics from multiple dimensions, providing a scientific basis for accurate teaching interventions [8]. The distinctive feature of the smart teaching environment is reflected in its ability to collect, analyze and apply educational data from multiple sources, which provides technical support for the accurate portrayal of learner profiles. For example, the learning engagement pattern of learners in Massive Open Online Courses (MOOCs) can be preliminarily outlined by analyzing their behavioral data such as video viewing duration, assignment submission frequency, and forum interaction content [9]. However, the prevalence of low learning engagement and insufficient learning effectiveness in smart teaching courses highlights the limitations of traditional teaching in terms of personalization and dynamic adaptability. How to realize precise teaching interventions based on learner persona has become a key proposition to promote the development of high-quality education.

To address the above background, this study proposes a learner persona model that integrates behavioral, cognitive and affective dimensions, and designs an accurate teaching intervention framework based on the Mahalanobis-enhanced K-Means++, outputs learner personas, and proposes accurate teaching intervention strategies for each group. Specifically, the research involves: (1) integrating multi-source data such as learning engagement, interaction, completion and emotional state to construct a learner labeling system; (2) optimizing the traditional K-means++ algorithm and introducing the Mahalanobis distance to solve the problem of multi-dimensional data correlation and improve the accuracy of subgroups; (3) conducting experiments on the Canvas open dataset and exploring ways to improve teaching effectiveness through targeted group intervention strategies.

The theoretical significance of this study lies in proposing a multidimensional user persona model to provide a more comprehensive analysis framework for accurate teaching. At the practical level, the Mahalanobis-enhanced K-Means++ and intervention strategy provide actionable solutions for personalized teaching in smart education scenarios, with a view to providing new theoretical and technical support for the digital transformation of education. The core contribution lies not in a novel clustering algorithm in itself, but in a systematic empirical analysis demonstrating the suitability of covariance-aware metrics for educational data and mapping the resultant clusters to actionable intervention strategies.

To guide our investigation, this study addresses the following research questions:

1. How can a multidimensional learner persona model, integrating behavioral, cognitive, and motivational expectations data, be constructed to effectively segment students in a smart learning environment?
2. In the context of clustering these multidimensional personas, does the Mahalanobis distance offer a significant performance improvement over the standard Euclidean distance within a K-Means++ framework, and what are the implications of the results for methodological choices in educational data mining?

2. Related Work

2.1. Learner Persona

Alan Cooper, the father of interaction design in the United States, first proposed the term "user persona", which he considered to be a target user model built on a series of real data [10]. Learner persona is the application of user persona in the education field, which is a visual learning analysis technology to reflect the behavioral characteristics and information of learners, aiming

at identifying and classifying learner types through the persona in order to provide appropriate teaching services for them. At the theoretical level, different researchers have analyzed the value and visualization of learner persona from different perspectives. The U.S. Instructional Management System Global Learning Consortium published the learner information model IMS Learner Information Package (IMS LIP), one of the most widely used international learner models, which adopts the XML standard data framework to support learner information with other similar specifications for recording and managing learner-related qualifications, goals and achievements [11].

Compared with theoretical research, scholars in recent years have paid more attention to the application functions of learner personas, including specific applications in the education field such as teaching agent, intelligent behavior inference, smart campus, and academic alert [12]. In the field of teaching agent application, Feng *et al.* [13] created detailed learner personas, integrating advanced generative pre-trained models such as GPT, fine-tuned with data related to English language teaching, aiming to innovate the English learning experience with AI technology and achieve personalized and efficient teaching guidance. Alshaikh and Hewahi [14] proposed a pedagogical model for Intelligent Tutoring Systems using a deep reinforcement learning architecture. This model delivers personalized learning experiences based on students' individual learning styles and preferences. Furthermore, they suggested that affective factors can significantly influence learning outcomes and advocated the use of learner personas to intelligently infer and adapt to student behaviors. Suleman *et al.* [15] introduced a conversational agent negotiation mechanism into an open learner model, which uses the learner's affective and behavioral states to intelligently infer student behaviors and purpose, increasing student engagement and reflection, thus improving the accuracy of the learner model. Maddu and Murugappan [16] proposed a novel student engagement prediction model for online learners based on facial emotion which monitors the engagement level of online learners in order to provide targeted instructional interventions.

Some scholars have applied learner personas to academic early warning. Xu [17] designed a visualization system for student behavior analysis to collect relevant information such as students' grades, attendance, book borrowing, *etc.*, analyze and mine the data, and present it with visualization tools to build an early warning management platform. Practice has shown that the visualization system can accurately analyze the academic situation, intelligently analyze and warn students of disciplinary violations, provide a reliable reference for student behavior management, and has high feasibility and reliability. Duong *et al.* [18] used the power of massive data sources from the education sector and the modernity of machine learning technology to construct an academic early warning system based on academic performance, which directly reflects students' academic status. In terms of smart campus applications, Zhang *et al.* [19] developed a college student persona system utilizing Hadoop big data technology. By combining the Canopy and k-means algorithms, student information is analyzed in multiple dimensions, which provides a way for schools to understand students more deeply.

2.2. Precision Teaching Interventions

In the 1960s, American scholar Ogden Lindsley, influenced by Skinner's behaviorist theory, proposed a teaching procedure with evidence-based attributes — "Precision Teaching" [20]. In recent years, with the wide application of big data, data mining and analysis and other technologies in education, education and teaching methods have undergone profound changes, and precision teaching and precision research have received increasing attention from researchers, and precision teaching interventions based on big data have been continuously verified in theory and practice [21-23].

Precision teaching has gone through a development process from being based on intuitive data to being based on a framework to being based on data analysis under the current development of intelligent technology, reflecting the trend of interaction between teachers and students in relation to educational data, as well as the fusion of instrumental attributes and value attributes of technology. The concept of humanism has gradually received attention in

the development of the connotation of accurate teaching, and the cognitive, emotional, behavioral and other data in the learning process of learners have become the basis for research. The ultimate goal is to promote teachers' accurate "teaching" and students' personalized "learning" [24, 25]. For example, Wang *et al.* [26] proposed a data-driven precision teaching intervention mechanism combining Educational Data Mining (EDM) and Learning Analytics (LA) technologies, which can provide early warnings and reminders to students at risk and provide targeted teaching help and support to students at different levels. Al-Murtadha [27] used an intervention study to help English language learners (ELLs) visualize themselves as possible successful second language speakers, and the results showed that appropriate instructional interventions facilitated the students' learning experience. Similarly, Gui *et al.* [28] assessed the impact of media education interventions on screen time management and expected use of mobile devices, emphasizing the effectiveness of incorporating media education interventions into daily instruction to promote student achievement and develop content-related digital skills.

Some scholars apply learner persona to precise teaching interventions, exploring learning behavior paths, teaching laws and teaching modes from the perspective of learner persona construction, which is a breakthrough and research focus of intelligent education and teaching. Diao *et al.* [29] proposed the E-GPPE-C model of intelligent learning by using artificial intelligence technology, which includes the learning environment, learning paths, learning evaluation, learner persona, education knowledge mapping and learning communities, and is used to explain the operating mechanism, elements and features of intelligent learning. Liu and Yuan [30] investigated how to construct a learner persona model in a blended learning environment, developed personalized teaching intervention strategies based on learner persona, and evaluated the effectiveness of these strategies. Jiang [31] introduced AI technology in English distance education to generate learner profiles to portray learning behaviors, which ensures accurate teaching for the personalized needs of English learners. Zhang [32], on the other hand, proposed the direction of constructing an intelligent technology-supported higher mathematics teaching model, modeling learner persona based on subject knowledge mapping of learning outcome data, searching for optimal learning paths using an improved ant colony optimization algorithm, and then combining with an improved convolutional neural network to generate personalized learning paths and teaching intervention strategies.

To situate our contribution within the existing literature, we highlight several key works and the specific gaps our study aims to address. For instance, the E-GPPE-C model proposed by Diao *et al.* provides a comprehensive conceptual framework for smart learning elements. However, this work remains theoretical, lacking the empirical validation and specific algorithmic implementation for persona clustering that our study provides. Similarly, Liu and Yuan demonstrate the practical application of personas for developing intervention strategies within a blended learning context, but their focus does not include a critical comparison of different clustering algorithms, which is a central methodological contribution of our paper.

Other studies have explored the use of AI in specific educational domains. Jiang highlights the utility of AI in personalizing English distance education, but the work lacks detail on the specific algorithms used and their performance evaluation. Zhang innovatively applies ant colony optimization to model learner personas for higher mathematics, yet the focus is on learning path optimization rather than the foundational clustering of personas, and it does not incorporate the affective dimensions that are integral to our multidimensional model.

By contrast, this study makes a distinct contribution by: (1) constructing a holistic, multidimensional persona model that integrates behavioral, cognitive, and motivational expectations states; and (2) providing a critical, data-driven comparative analysis of Mahalanobis-enhanced K-Means++ clustering performance with both Euclidean and Mahalanobis distance metrics on a public dataset. This approach addresses the common focus on single-dimensional data and the need for a more rigorous evaluation of clustering methodologies in the field of precision teaching (Table 1).

Table 1. Comparative analysis of related works in learner persona and precision teaching.

Author (Ref)	Core Methodology	Data Dimensions	Application Context	Key Limitations Addressed by Current Study
Feng *et al*. [13]	Generative Pre-trained Models (GPT)	Textual & Linguistic Data	English Language Tutoring	Focuses on content generation; lacks structural clustering of learner archetypes.
Alshaikh and Hewahi [14]	Deep Reinforcement Learning (DRL)	Learning Styles & Preferences	Intelligent Tutoring Systems (ITS)	High computational complexity; lacks explicit consideration of multidimensional feature correlations.
Maddu and Murugappan [16]	Facial Emotion Recognition (FER)	Real-time motivational expectations states	Online Video Learning	Relies solely on physiological/ affective signals; ignores longitudinal behavioral patterns in LMS.
Zhang *et al.* [19]	Canopy + K-Means (Standard)	Behavioral & Academic Records	Smart Campus Management	Uses standard Euclidean distance; fails to account for covariance between behavioral features (*e.g.*, time vs. grade).
Liu and Yuan [30]	Blended Learning Profiling	Course Completion Data	Blended Learning	Single-dimensional grouping; lacks dynamic distance metric optimization for precise segmentation.
Current Study	Mahalanobis--Enhanced K-Means++	Behavioral, Cognitive, & Affective	Canvas MOOC Environment	Integrates multidimensional holistic methodology; optimizes clustering accuracy via covariance-aware distance metric.

2.3. Literature review

In summary, many scholars have affirmed the role of visualization of learner persona and its important value in the field of education and teaching, and the research focuses on the practical level of application, attaching importance to the functional attributes as well as the technical algorithms, *etc.*, but there are still fewer applications in smart teaching scenarios. In addition, studies have shown that learner personas provide a breakthrough for the realization of precision teaching, which helps to realize personalized learning and precision teaching, and many scholars take learner persona as a logical starting point in the implementation of precision teaching, but studies focus on a single dimension, ignoring the dynamic correlation between cognitive level and motivational expectations state, which makes it difficult to meet the demand for precision intervention. Although the K-means clustering method is widely used for learner clustering, its traditional algorithm relying on Euclidean distance has difficulty effectively dealing with the complex associations of multi-dimensional data, which restricts the relevance and real-time performance of intervention strategies. Therefore, how to construct a comprehensive and reasonable learner profiling model and accurately guide the implementation of teaching and learning is still an urgent problem to be solved.

3. Model Construction

3.1. Learner persona construction process

The study follows the overall process of instructional design and proposes the process of constructing a learner persona in a smart teaching environment, as shown in Figure 1. The overall process is learner-centered and consists of three key components: first, collecting multivariate data and performing data cleansing, data transformation and data mining on it; second, constructing a learner persona model and generating a learner persona; and finally, generating a teaching intervention strategy based on the learner persona, which can be applied to real teaching scenarios.

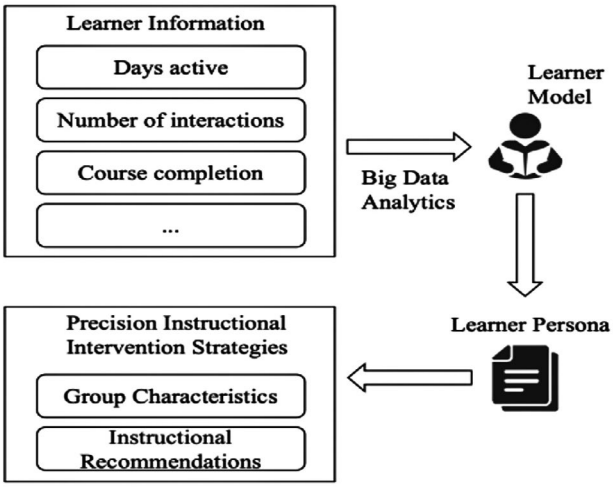


Figure 1. Learner persona construction process.

3.2. Modeling learner persona

Based on the relevant domestic and international literature, the study focuses on the behavior, cognition and emotion of learners, reflecting "whether to learn", "how to learn" and "why to learn", respectively, in the context of a smart teaching environment. We constructed a learner persona model that includes three dimensions: behavioral characteristics, cognitive level, and motivational expectations state, and identified individual and group characteristics of learners through the persona to help teachers implement personalized teaching interventions, as shown in Figure 2.

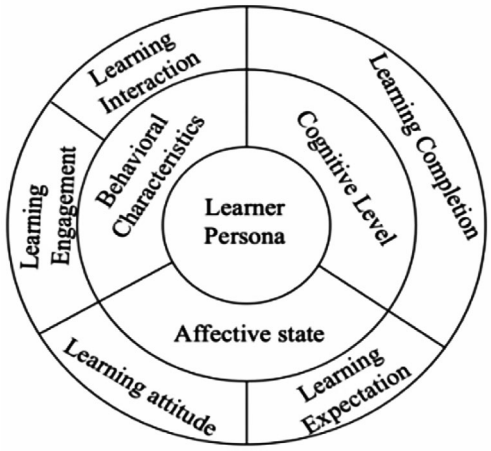


Figure 2. Learner persona model

(a) Behavioral characteristics

Behaviorist theory emphasizes observable behavior as a direct reflection of learning and considers learning as a stimulus-response linkage process [33]. Behavioral characteristics reflect the collection of characteristics of learners involved in learning activities, and learners' behavioral inputs can directly reflect the learning status of learners, and the analysis of learners' learning behavioral characteristics and patterns can be of great help in realizing accurate teaching. The division of behavioral characteristics dimensions varies according to the research scenarios, objects and conditions. Based on the analysis of the literature [34] and the data that can be collected, this study includes two sub-dimensions of learners' behavioral characteristics: learning engagement and learning interaction, where learning engagement mainly considers the number of active days and learning interaction mainly considers the number of interactions.

(b) Cognitive level

Knowledge is acquired through cognition, which is related to learners' understanding, thinking, problem solving and information processing [35-37], and the cognitive level mainly refers

to the core competencies that can be mastered in the process of learning subject knowledge, which is directly related to the achievement of teaching objectives. The mastery learning theory proposed by Bloom defines learning attainment as the ratio of the actual learning time to the necessary learning time, emphasizing that the progress of completing learning tasks is the basis of cognitive development. Therefore, the sub-dimension of cognitive level in this study is learning attainment.

(c) Motivational expectations state

According to psychologist Reinhard Pekrun, learning affect is a variety of emotional experiences felt by students in learning or teaching situations, including excitement, anxiety, curiosity, satisfaction and various other emotional states, and learning affect runs through the entire learning stage, not only affecting students' learning efficiency and motivation, but also having an impact on their engagement, attention and thinking in the classroom, and studies have shown that positive learning affect can significantly enhance learning effects. In this study, due to the limitations of the Canvas dataset, direct physiological or real-time emotional signals are not available. The "Motivational expectations" dimension is therefore operationalized using self-reported proxy variables available within the LMS, specifically learning attitudes and learning expectations. Therefore, affective learning is an extremely important educational issue in educational scenarios and is significant in enhancing learners' learning experience and educational outcomes. In addition, self-determination theory states that motivation for learning promotes academic success, psychological well-being, and long-term personal development. Therefore, this study focuses on learners' motivational expectations state and divides it into two sub-dimensions: learning attitudes and learning expectations.

3.3. Data preprocessing and implementation details

To ensure the validity and reproducibility of our clustering experiments, a systematic data preprocessing pipeline was implemented.

Data elimination: The missing value cleaning of the Canvas dataset is performed on the raw records, and abnormal records with a missing value ratio exceeding 30% are removed.

Handling of missing values: The Canvas dataset contained a small percentage of missing values in numerical columns. These were handled using mean imputation, where each missing value was replaced with the mean of its respective feature column.

Handling of categorical variables: For the categorical feature Learning objectives in the learning goal dimension and User type, one-hot encoding is first applied to convert the categorical values into numerical vectors. Subsequently, Principal Component Analysis (PCA) is employed to project the high-dimensional sparse vectors into a one-dimensional dense space, which eliminates the dimensionality explosion and sparsity issues while retaining the core feature information.

Sample aggregation: For students enrolled in multiple courses, global aggregation is conducted based on student ID. The mean value is adopted for "active days" and "interaction times", while the weighted mean is used for "expected weekly learning time" and "course completion rate".

Feature normalization: As the features in our persona model (*e.g.*, number of active days, course completion rate) exist on different scales, all numerical features were normalized to the range [0, 1] using Min-Max scaling. This prevents features with larger scales from disproportionately influencing the distance calculations, which is crucial for both Euclidean and Mahalanobis-based clustering.

Through the above process, 325,000 raw records in the Canvas dataset were aggregated into 2,478 student-level samples via the proposed processing pipeline.

Implementation: All experiments were conducted using Python 3.9 and the scikit-learn library (version 1.2.2). The implementations of K-Means++ and DBSCAN were sourced directly from this library. For the Mahalanobis distance calculation within K-Means++, the sample covariance matrix was computed for

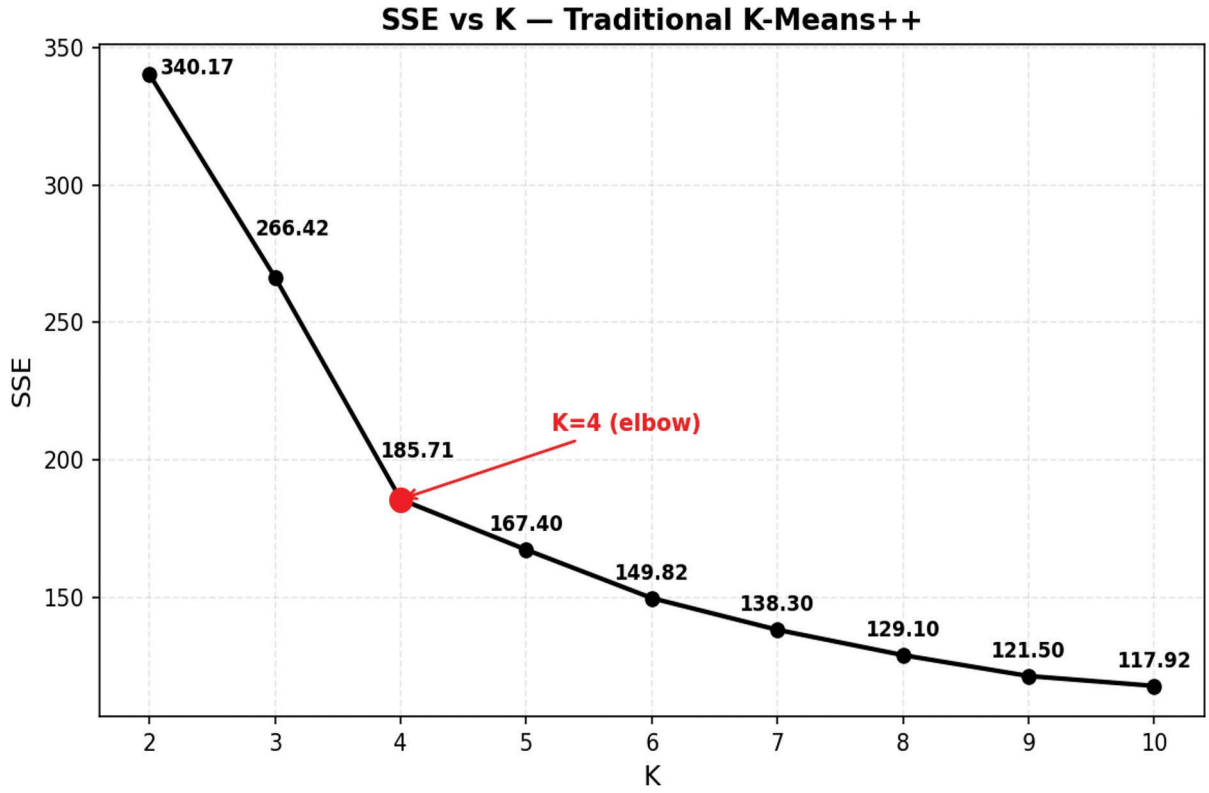


Figure 3. Elbow method for optimal K (K-Means++).

the features designated in Table 3. The number of clusters (k) for K-Means++ was set to 4 based on preliminary analysis using the elbow method (as illustrated in Figure 3, which shows the SSE curve with a clear elbow at $k = 4$) and domain knowledge. For DBSCAN, the parameters eps and min_samples were optimized using a grid search approach to maximize the Silhouette Score.

3.4. Mahalanobis-enhanced K-Means++

Labels, as the core of learner profiling, can be a generalization of single or multiple features. In order to be able to better understand the learner's learning situation, this paper combines multiple competency features mined in Section 3.2 and uses clustering to divide learners into groups, summarizes the learner's learning situation by analyzing the final division results, and implements personalized teaching interventions.

In this paper, we use the K-Means++ algorithm for clustering, an unsupervised learning method that improves on the K-Means algorithm. Although the traditional K-Means algorithm is easy to understand and implement, and is fast with good clustering performance, it has some shortcomings: (1) the number of clusters K must be set in advance, and the choice of K has a great impact on the clustering results; (2) K-Means depends strongly on the initial values: because the initial cluster centers are selected randomly, the clustering results are prone to differ from run to run. The first deficiency cannot be easily resolved by K-Means or its derivative algorithms. Usually, the choice of K value can be determined according to the content of the problem, and in the case of uncertainty, the choice is based on some methods, such as the "elbow method". The second deficiency is effectively addressed by the K-Means++ algorithm. The steps of the K-Means++ algorithm are shown in Table 2.

The K-means++ algorithm is improved as follows:

The degree of distortion of each class in clustering reflects the quality of the clustering, and the degree of distortion is the sum of the squares of the distances from each sample to the cen-

Table 2. Steps of K-means++ algorithm.

1.	A randomly selected point in the data set to be input is used as the clustering center μ.
2.	For the dataset, calculate the distance D between each point x and the nearest cluster center already selected.
3.	Select a new data point as a clustering center based on the distance D calculated in step 2 (the larger the value of D the higher the probability of being selected as a clustering center).
4.	Repeat steps 2 and 3 until k cluster centers have been selected.
5.	The k centers obtained are used as the initial centers, and the standard K-Means algorithm is run.

ter of mass position under each class. In the K-Means++ algorithm, the sum of the squares of the distances from each sample to the center of mass is also the main basis for classification, so the calculation method of the distance from the sample to the center of mass has a non-negligible impact on the final clustering effect, and it is necessary to reconfigure the center of mass formula according to the characteristics of the data.

The traditional K-means++ algorithm generally uses the Euclidean distance to measure the distance between samples. The Euclidean distance is mainly considered as a numerical distance. Euclidean distance is defined as follows:

In an *N*-dimensional space, the Euclidean distance between two points is obtained by subtracting their coordinates in each dimension, summing the squared differences, and taking the square root of the sum.

Euclidean distance is used very widely. For example, in one dimension, the distance between two points measured along the coordinate axis is their Euclidean distance. In the two-dimensional case, the Euclidean distance is obtained by subtracting the coordinates of the two points in each dimension, squaring the differences, summing them, and taking the square root, that is:

$$c = \sqrt{\left(x_1 - x_2\right)^2 + \left(y_1 - y_2\right)^2}$$

In the three-dimensional case, the Euclidean distance corresponds to the length of the diagonal of a rectangular box whose three side lengths are known: the differences between the coordinates of the two points in each of the three dimensions are squared and summed, and the square root is taken, that is:

$$c = \sqrt{\left(x_1 - x_2\right)^2 + \left(y_1 - y_2\right)^2 + \left(z_1 - z_2\right)^2}$$

Mahalanobis distance focuses on the performance of the data in terms of the covariance distance, which can be regarded as a generalization of the Euclidean distance and overcomes the shortcomings of the Euclidean distance when evaluating dimensions that are not scaled uniformly and are correlated. That is, Euclidean distance only calculates numerical distances and ignores the dependence between data attributes, which is compensated by the Mahalanobis distance.

The mean and covariance are essential for computing the Mahalanobis distance. For the mean vector

$$\mu = \left(\mu_1, \mu_2, \mu_3, ..., \mu_p\right)^T$$

and a multivariate vector with covariance matrix Σ,

$$x = \left(x_1, x_2, x_3, ..., x_p\right)^T$$

Then the Mahalanobis distance is:

$$D_M(x) = \sqrt{\left(x - \mu\right)^T \Sigma^{-1} \left(x - \mu\right)}$$

The Mahalanobis distance can also be defined as the degree of difference between two random

variables X and Y with covariance matrix Σ. The Mahalanobis distance between data points x, y is:

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \Sigma^{-1} (\vec{x}-\vec{y})}$$

Where Σ is the covariance matrix of the multidimensional random variable and μ is the sample mean. In this study, the covariance matrix Σ is computed globally based on the entire standardized dataset to capture the overall correlation structure between the learning expectation features (learning objectives and expected weekly learning time).

In this paper, according to the analysis of the data characteristics of the Canvas dataset and the labeling architecture involved in this paper, it is found that there are multiple data interacting with each other in some complex labels, and in order to better divide the clusters in the clustering, it is intended to replace the traditional Euclidean distance in the K-means++ algorithm in the processing of this part of the data by using the Mahalanobis distance, while for basic labels the Euclidean distance is sufficient, as shown in Table 3. When calculating the Mahalanobis distance for the two labels within the Learning Expectations dimension, a robust covariance estimator was adopted. By selectively identifying the core valid subset of samples in the dataset, outliers were effectively mitigated, thereby resolving the issue of distortion in the covariance matrix typically caused by extreme values (*e.g.*, anomalous behavioral data). Simultaneously, to address the potential singularity of the covariance matrix, a regularization term $\Sigma_{reg} = \Sigma + \lambda I$ (where $\lambda = 10^{-6}$) was incorporated to ensure matrix invertibility, further enhancing the computational stability, where Σ is the covariance matrix of the multidimensional random variable.

Experiments are also repeated across 10 to 20 random seed configurations, and the results further verify the stable and superior performance of the proposed algorithm.

During clustering, based on the k centroids obtained with the procedure in Table 2 and the defined Mahalanobis distance formula, the features of each sample x_i are divided into an independent feature set and a correlated feature set (learning expectation):

1. Calculate the Euclidean distance from sample x_i to each centroid c_j on the independent feature set.
2. Calculate the Mahalanobis distance from sample x_i to each centroid c_j on the correlated feature set, where a robust covariance matrix is adopted.
3. Fuse the two types of distances to obtain the combined distance, which is the sum of the Euclidean distance and the Mahalanobis distance, and assign sample x_i to the cluster corresponding to the centroid with the minimum combined distance.

Table 3. Mahalanobis-enhanced K-Means++ distance calculation method.

Labels	Multiple raw data	Is there a correlation	Distance calculation method
Learning Engagement	no	no	Euclidean distance
Learning Interactivity	no	no	Euclidean distance
Learning Completion	no	no	Euclidean distance
Learning Attitudes	no	no	Euclidean distance
Learning Expectations	yes	yes	Mahalanobis distance

4. Experimental Results

4.1. Data acquisition and analysis

The first step in learner profiling is to obtain accurate, reliable and comprehensive data, and there are two ways to obtain such data: researchers either collect their own data or use open datasets published by international organizations or institutions. Self-collected data can come from system logs, which provide large volumes of user behavior data but are generally difficult to obtain and require extensive cleaning; alternatively, user data can be collected with web crawlers (*e.g.*, written in Python), but such data are generally less complete and may be skewed to some degree. The second method is to download the data directly from major organizations, but in many cases user behavioral data contains private information about the user, and in order to protect the user's privacy and security these organizations need to carry out a series of anonymization procedures on the data before the final release. Nowadays, a growing number of open datasets have been published both domestically and internationally.

The dataset selected for this paper is the Canvas open dataset. The Canvas open dataset contains more than 325,000 user behavior records covering the period from January 2014 to September 2015. These records reflect the performance of user activities in 238 courses. The dataset includes many features that record basic information about the user and information about their behavior in a particular course as well as information about the course.

Based on the above data analysis and literature summarization, the persona labeling system is summarized as shown in Table 4.

4.2. Algorithm comparison

Many indices can be used to evaluate clustering. This paper selects three of them to measure and compare the clustering results: the Silhouette Score, the Davies–Bouldin Index (DBI), and the Calinski–Harabasz Index (CH).

The traditional K-Means++ and the Mahalanobis-enhanced K-Means++ are compared. When calculating the learning expectation, because it involves multiple original data and there is a certain correlation between the data, the centroid distance should be calculated with the Mahalanobis distance. Therefore, for the learning expectation label, this experiment calculates the centroid distance using the Euclidean distance and the Mahalanobis distance, respectively, and then performs clustering and calculates the relevant evaluation indices.

First, the clustering results obtained by the K-Means++ algorithm using the Euclidean and Mahalanobis distances are presented. Figure 4 shows the clustering results when the Euclidean distance is used to calculate the centroid dis-

Table 4. Persona label system.

One-dimensional	Two-dimensional	Data label
Behavioral characteristics	Learning Engagement	Active days
	Learning Interactivity	Number of interactions
Cognitive level	Learning Completion	Course completion rate
Motivational expectations state	Learning Attitudes	User type
	Learning Expectations	Learning objectives, expected weekly learning time

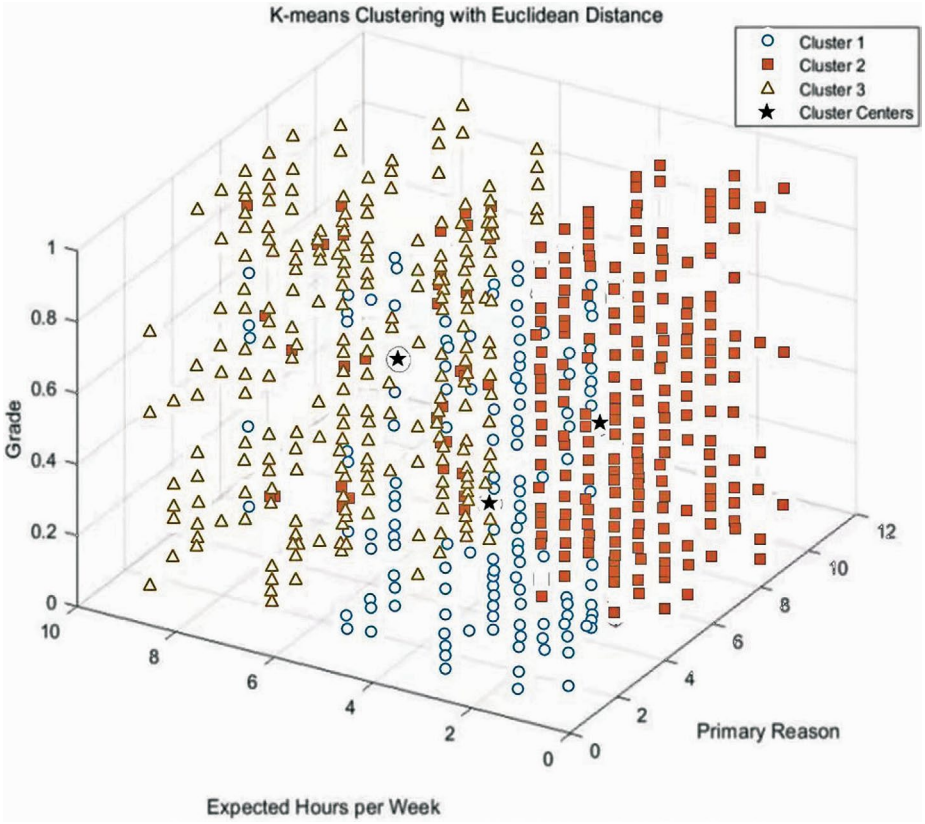


Figure 4. Learning expectation label Euclidean distance clustering effect.

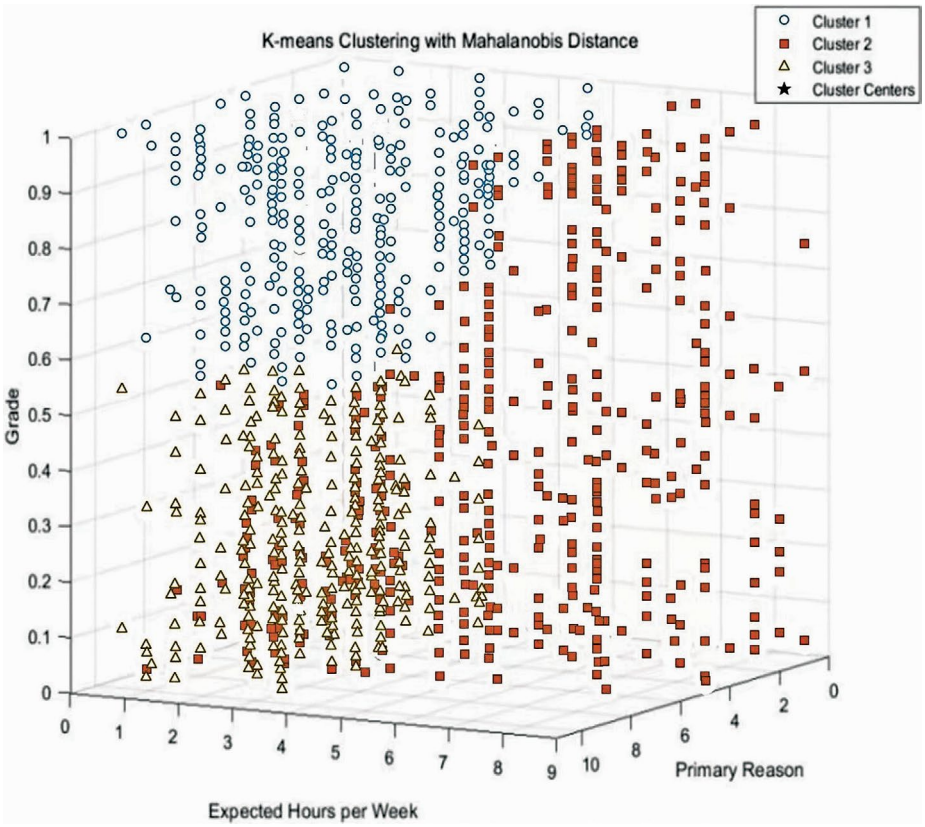


Figure 5. Learning expectation label Mahalanobis distance clustering effect.

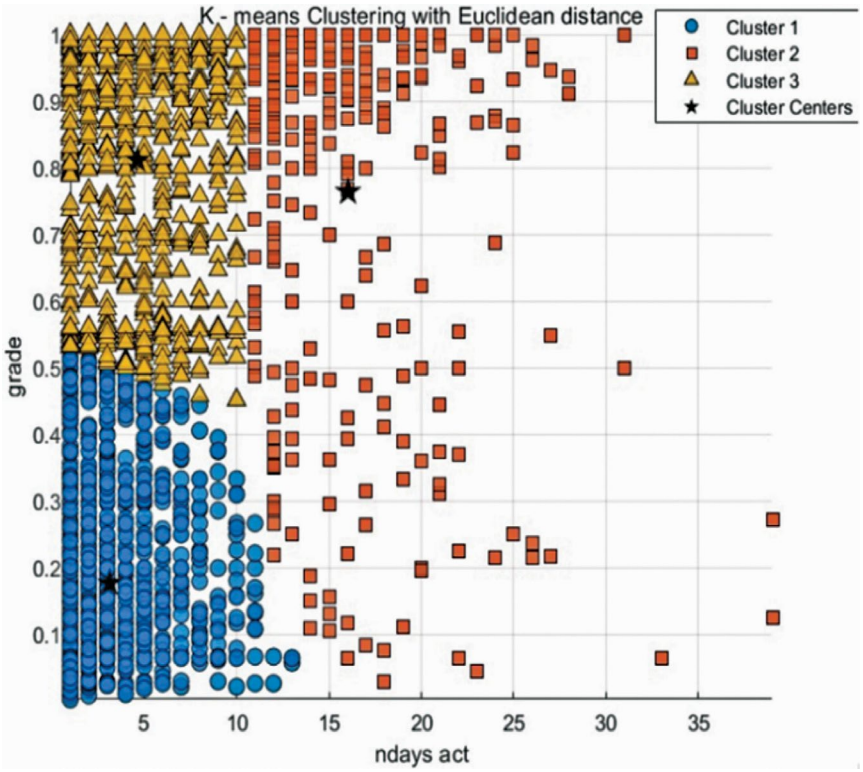


Figure 6. Learning engagement label clustering.

tance. In Figures 4, 5, and 10, the axis labeled 'Primary Reason' represents the categorical feature encoded from the 'learning objectives' data (*e.g.*, obtaining a degree, personal interest), while 'Grade' and 'Expected Hours per Week' represent normalized numerical values.

Figure 6 shows the clustering results for the learning engagement label as the center of mass calculation. The results for the three indicators are shown in Table 5.

The results for the three indicators are shown in Table 6. Comparing the clustering effect of using Euclidean distance and using Mahalanobis distance, the results of their three metrics are shown in Table 7.

The superior performance of the K-Means++ algorithm when using Mahalanobis distance, as evidenced by improvements across all three evaluation metrics (Silhouette Score: 0.4277 vs. 0.3281; Davies-Bouldin Index: 0.8118 vs. 0.8279; Calinski-Harabasz Index: 1906.5596

Table 5. Learning expectation label Euclidean distance clustering evaluation metrics.

Indicator	Methods of judging effectiveness	Value of the indicator
Silhouette Score	The bigger it is, the better it works.	0.3281
Davies-Bouldin	The smaller it is, the better it works.	0.8279
Calinski-Harabasz Index	The bigger it is, the better it works.	1472.1310

Table 6. Learning expectation label Mahalanobis distance clustering evaluation metrics.

Indicator	Methods of judging effectiveness	Value of the indicator
Silhouette Score	The bigger it is, the better it works.	0.4277
Davies-Bouldin	The smaller it is, the better it works.	0.8118
Calinski-Harabasz Index	The bigger it is, the better it works.	1906.5596

Table 7. Summary of comparative clustering algorithm performance metrics.

Clustering Algorithm	Distance Metric / Methodology	Silhouette Score (↑)	Davies-Bouldin Index (↓)	Calinski-Harabasz Index (↑)	Running time (s)	Performance Interpretation
Improved K-Means++ (Proposed)	Mahalanobis Distance	0.4277	0.8118	1906.56	0.2395 s	Optimal: Effectively handles feature correlation via covariance matrix.
Traditional K-Means++	Euclidean Distance	0.3281	0.8279	1472.13	0.0866 s	Suboptimal: Assumes spherical clusters, failing to capture elliptical learner distributions.
DBSCAN	Density-Based	0.1532	1.5594	1288.16	0.0975 s	Poor: Dataset lacks distinct density separation; globular cluster assumption holds stronger validity.

vs. 1472.1310), is not merely a numerical artifact but a direct consequence of the data's underlying structure. This finding is particularly significant for the 'Learning Expectations' label, which was constructed from two correlated features: 'learning objectives' and 'expected weekly learning time'. Standard Euclidean distance treats these features as independent axes in the feature space, failing to account for their covariance. In contrast, the Mahalanobis distance incorporates the covariance matrix into its calculation, effectively rescaling the feature space to handle inter-variable correlations. This makes it exceptionally well-suited for multi-dimensional educational data, where learner attributes are rarely orthogonal. The higher Silhouette Score indicates that the resulting clusters are denser and more clearly separated, leading to more distinct and interpretable learner personas. The computational cost of the proposed algorithm is proportionally higher: the median runtime on the same dataset is 0.2395 s, approximately 2.77× that of the traditional K-Means++ (0.0866 s). This overhead is attributable to the inversion of the covariance matrix required for Mahalanobis distance computation at each K-Means iteration, which, however, remains well within practical limits for datasets of this scale.

To further validate our choice of the K-Means++ framework, we compared its performance

against a widely used density-based clustering algorithm, DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Unlike K-Means, DBSCAN does not require the number of clusters to be specified in advance and can identify arbitrarily shaped clusters. We performed a grid search to optimize DBSCAN's hyperparameters, eps (the maximum distance between two samples for one to be considered as in the neighborhood of the other) and min_samples (the number of samples in a neighborhood for a point to be considered as a core point), to maximize the Silhouette Score on our dataset. The noise samples generated during the clustering process are eliminated and excluded from the final clustering results.

The optimal DBSCAN configuration yielded a Silhouette Score of 0.1532, a Davies-Bouldin Index of 1.5594, and a Calinski-Harabasz Index of 1288.16. Noise samples accounted for 7.1% of the data. Compared to both K-Means++ variants, DBSCAN exhibits a substantially lower Silhouette Score, a higher Davies-Bouldin Index, and a lower Calinski-Harabasz Index. These results indicate that the underlying cluster structure in the Canvas dataset is more globular and less defined by varying densities, making centroid-based approaches like K-Means++ better suited to capture compact, well-separated clusters. This cross-metric comparison reinforces the selection of the proposed Improved K-Means++ algorithm as the foundational method for our persona modeling. In terms of computational efficiency, the proposed algorithm requires 0.2395 s per run, approximately 2.5× that of DBSCAN (0.0975 s), reflecting the additional overhead of Mahalanobis distance computation per iteration. Nonetheless, this cost remains negligible in absolute terms (<0.25 s) and is well within practical deployment constraints for datasets of this scale.

4.3. Learner persona output

Based on the above results, this paper uses the Mahalanobis-enhanced K-Means++ to cluster the original data in the dataset according to the labeling system; then, according to the formal description of the labels, the original data are converted into labeled data, which accurately translates the clustering results into the user persona data needed in this paper.

(a) Learning Engagement Label

Learning engagement is mainly measured by the number of active days; these data were clustered with the following results.

By integrating the clustered data, the following label distributions were obtained, as shown in Figure 6 and Table 8.

In this paper, learning engagement is pre-described as: Learning Engagement = {High, Medium, Low}. By analyzing the number of labels and the data after clustering, it is concluded that label 0 learning engagement is low, corresponding to the blue point, label 1 learning engagement is medium, corresponding to the yellow point, and label 2 learning engagement is high, corresponding to the red point.

Table 8. Learning engagement label statistics.

Labels	**0**	**1**	**2**
Amount	1221	940	317

According to the combination of three parts of data: the labeled data after clustering and the preliminary analysis data as well as the original data, this paper changes the original dataset into the data needed in the new learner persona model, creates a new learning engagement column, sets the label 0 learning engagement to low, label 1 learning engagement to medium, and label 2 learning engagement to high, and replaces the original number of active days column, which makes the data more reasonable and allows administrators to directly view a user's learning engagement. The resulting labels demonstrate good interpretability.

(b) Learning interactivity label

Learning interactivity is mainly measured by the number of interactions; these data were clustered with the following results:

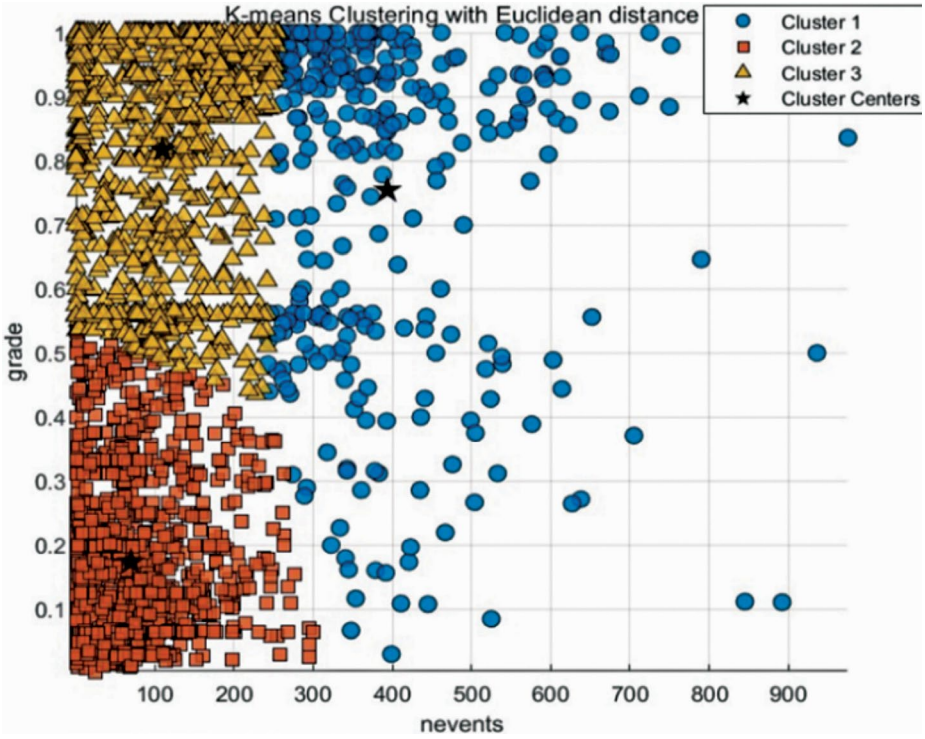


Figure 7. Learning interactivity label clustering.

By integrating the clustered data, the following label distributions were obtained, as shown in Figure 7 and Table 9.

Table 9. Learning interactivity label statistics.

Labels	**0**	**1**	**2**
Amount	1218	934	326

In this paper, the learning interactivity is pre-described as: Learning Interactivity = {High, Medium, Low}. By analyzing the number of labels and the data after clustering, it is concluded that label 0 learning interactivity is low, corresponding to the red point, label 1 learning interactivity is medium, corresponding to the yellow point, and label 2 learning interactivity is high, corresponding to the blue point.

According to the combination of three parts of data: the labeled data after clustering and the preliminary analysis data as well as the original data, this paper changes the original dataset into the data needed in the new learner persona model, creates a new learning interaction degree column, sets the label 0 learning interaction degree as low, the label 1 learning interaction degree as medium, and the label 2 learning interaction degree as high, and replaces the original interaction number column, which makes the data more reasonable and allows administrators to directly view a user's learning interactivity. The resulting labels demonstrate good interpretability.

(c) Learning completion label

Learning completion is mainly measured by the course completion rate; these data were clustered with the following results:

By integrating the clustered data, the following label distributions were obtained, as shown in Figure 8 and Table 10.

Table 10. Learning interactivity label statistics.

Labels	**0**	**1**	**2**
Amount	1282	1119	77

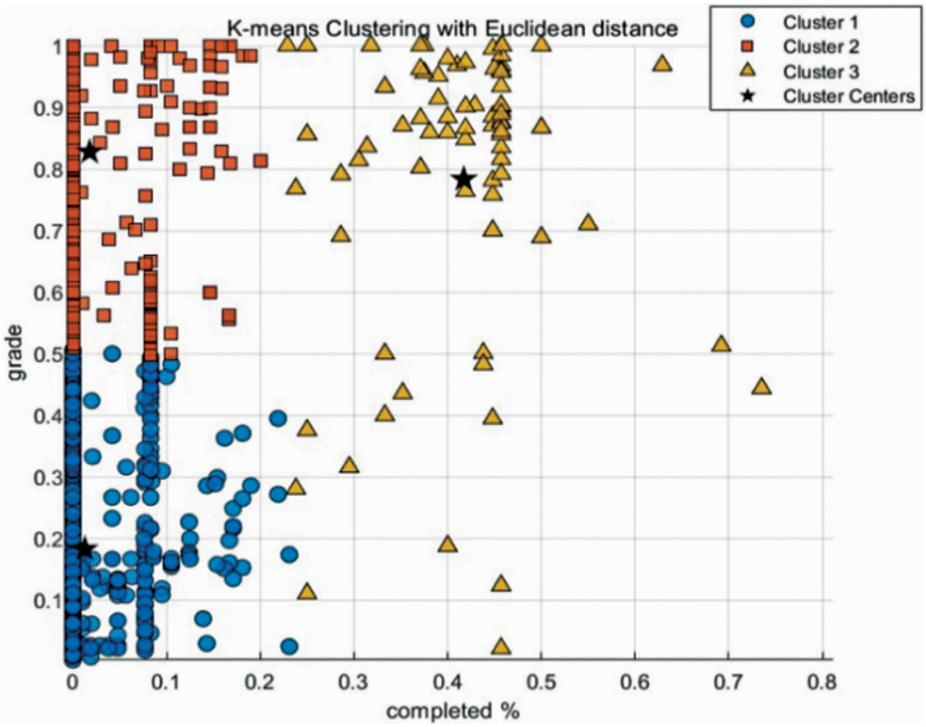


Figure 8. Learning completion label clustering.

In this paper, learning completion is pre-described as: Learning Completion = {High, Medium, Low}. By analyzing the number of labels and the data after clustering, it is concluded that label 0 learning completion is low, corresponding to the blue point, label 1 learning completion is medium, corresponding to the red point, and label 2 learning completion is high, corresponding to the yellow point.

According to the combination of three parts of data: the labeled data after clustering and the preliminary analysis data as well as the original data, this paper changes the original dataset to the data needed in the new learner persona model, creates a new learning completion column, sets label 0 learning completion to low, label 1 learning completion to medium, and label 2 learning completion to high, replacing the original course completion rate, which makes the data more reasonable and allows administrators to directly view a user's learning completion. The resulting labels demonstrate good interpretability.

(d) Learning attitude label

Learning attitude is mainly measured by user type; these data were clustered with the following results:

By integrating the clustered data, the following label distributions were obtained, as shown in Figure 9 and Table 11.

Table 11. Learning interactivity label statistics.

Labels	**0**	**1**	**2**
Amount	796	923	759

In this paper, learning attitude is pre-described as: Learning Attitude = {Active, Passive, Negative}. By analyzing the number of labels and the data after clustering, it is concluded that label 0 learning attitude is Active, label 1 learning attitude is Passive, and label 2 learning attitude is Negative.

According to the combination of the labeled data after clustering and the preliminary analysis data and the original data, this paper changes the original data set into the data needed in the new learner persona model, creates a new learning attitude column, sets label 0 learning attitude as active, label 1 learning attitude as

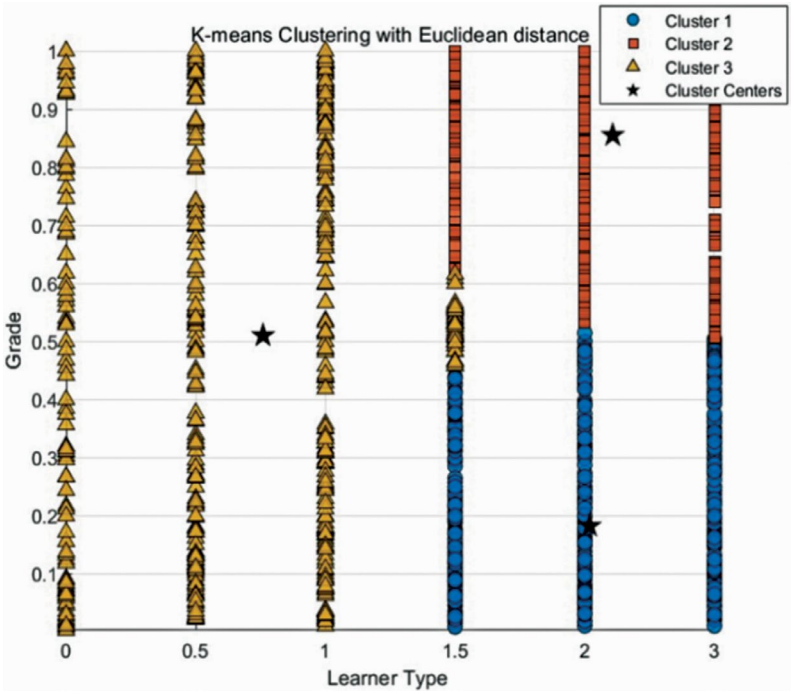


Figure 9. Learning attitude label clustering.

passive, and label 2 learning attitude as negative, replacing the original user type column, making the data more interpretable and convenient for the instructors and platform managers to directly analyze the learning attitude of a user. The resulting labels demonstrate good interpretability.

(e) Learning expectations label

Learning expectations are mainly measured by two raw variables, learning purpose and expected weekly learning time; these data were clustered, and the results are shown in Figure 10 and Table 12.

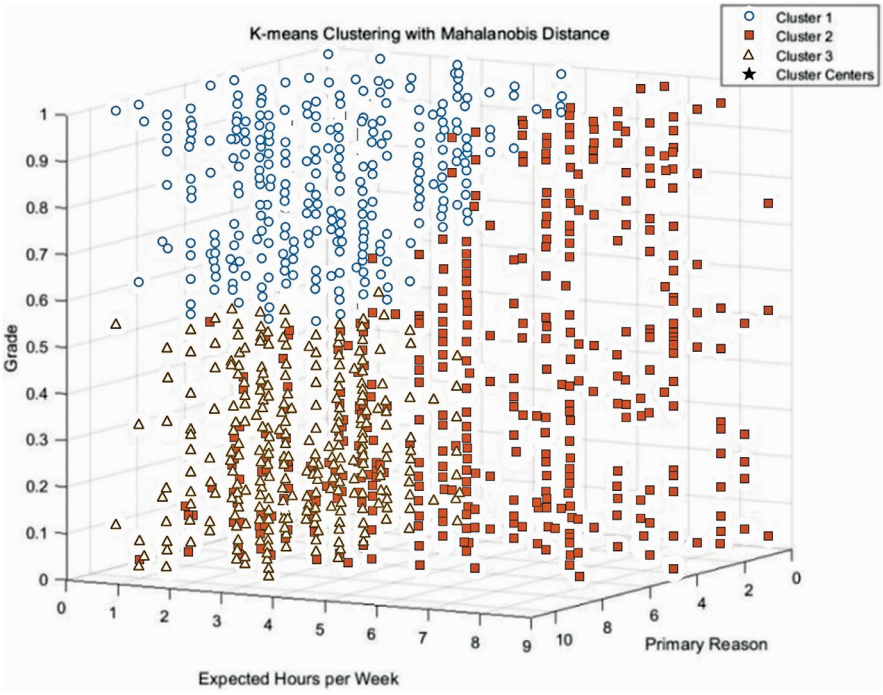


Figure 10. Learning expectations label clustering.

Table 12. Learning expectations label statistics.

Labels	0	1	2
Amount	952	901	625

In this paper, learning expectation is described as: Learning Expectation = {High, Medium, Low}. To maintain consistency with the labeling logic of other dimensions (*e.g.*, Learning Engagement, Learning Interactivity), a sequential labeling scheme is adopted for learning expectations. Under the clustering based on the enhanced algorithm in this section, the original data (weekly expected learning time and learning purpose) are integrated into a single labeled learning expectation. By analyzing the clustered data and clustering effects in Figure 10, the classification rules are determined as follows: label 0 for low learning expectation, label 1 for medium learning expectation, and label 2 for high learning expectation.

According to the combination of three parts of data (labeled data after clustering, preliminary analysis data, and original data), the original dataset is transformed into the data required by the new learner persona model. A new "Learning Expectations" column is created, where label 0 is set to low learning expectation, label 1 to medium learning expectation, and label 2 to high learning expectation. This replaces the original two columns of "weekly expected learning time" and "learning purpose", enabling administrators to directly analyze a user's learning expectation level with clear and consistent logic. The resulting labels demonstrate good interpretability.

(f) Classification of groups

A group lies between the whole class and the individual learner: it is a smaller category composed of learners with similar characteristics, and a class of students is generally divided into several such groups. Subdividing a class of students into subgroups allows for more precise differentiation of students and allows for more targeted instructional interventions.

The cluster persona mainly presents the level and characteristics of different learner groups, in which the style characteristics are static indicators, so this kind of data is used as a separate criterion for clustering. The data indicators of behavioral characteristics and ability level are selected for clustering, and in order to ensure the consistency of the data standard, the five indicators involved in clustering are firstly normalized by Min-Max, and the mean and standard deviation of the data of each of the 2478 students are obtained. After algorithmic clustering, the learners were divided into four category groups, as shown in Table 13.

Table 13. Mean values of category personas on each indicator.

	Cognitive level	Motivational expectations state		Behavioral characteristics	
	Learning completion rate	Learning Attitudes	Learning Expectations	Learning Interactivity	Learning Engagement
Group1	0.4168	0.7829	0.2829	0.1155	0.1748
Group2	0.0146	0.1526	0.8529	0.0202	0.0408
Group3	0.0217	0.6541	0.3233	0.0692	0.2807
Group4	0.0160	0.8382	0.0944	0.0314	0.0645

4.4. Instructional intervention strategies based on learner personas

Based on the above group classification and the labels assigned to each group, Group 1 has a positive attitude towards learning and the highest degree of learning completion, which is defined as "potential self-discipline learners"; Group 2 has the highest learning expectations, but the degree of learning completion and engagement is low, which is characterized by a disconnect between ideals and actions, which is defined as "high expectations of low-effective learners"; Group 3 has the highest degree of engagement in learning and a moderate learning attitude, which is defined as "active and participatory learners"; Group 4 has a positive learning attitude, but the degree of engagement and learning completion is low, which indicates that the initial enthusiasm fades quickly and falls into deep burnout, which is defined as "negative burnout type learners". The group labels, characteristics and problems of each cluster are shown in Table 14.

Detailed elaboration on each group's problems:

Group 1 (Potential Self-Discipline): Although they successfully complete courses, their reluctance to engage in interactive activities limits the transition from passive knowledge acquisition to active competence enhancement.

Group 2 (High Expectation Inefficiency): This group exhibits a significant intention-behavior gap. Their ambitious goals are undermined by poor time management and task avoidance, trapping them in a cycle of "fantasy and procrastination."

Group 3 (Active Participation): Their high activity level is often undirected. Without clear learning strategies, their frequent participation does not translate into effective knowledge accumulation.

Group 4 (Negative Burnout): Despite an initial positive attitude, these learners lack stress resilience. Without external support or flexible learning paths, their self-efficacy continuously erodes, leading to a state of deep burnout.

Precision teaching intervention is mainly based on the performance of individual learners in the learning process, using learner personas to present the learning characteristics of learners, judging the current learning status of learners, accurately diagnosing learning problems, and executing precision teaching intervention plans accordingly. Combining the 3P teaching model and social cognitive theory, this study proposes the following teaching intervention strategies in terms of task strategy, activity strategy, resource strategy, and evaluation strategy.

Table 14. Mean values of category personas on each indicator.

Group	Type of learner	Core features	Problems
Group1	Potential self-discipline learners	Positive learning attitude, highest learning completion	Limited interaction restricts application and practical expansion.
Group2	High expectations of low-effective learners	High learning expectations, low learning completion and engagement	"High expectation-low action" cycle; weak execution and procrastination.
Group3	Active and participatory learners	Highest engagement, moderate learning attitude, low learning completion	Inefficient participation; lacks goal-orientation and deep learning.
Group4	Negative burnout learners	Initial positive attitude; subsequent engagement and completion plummet	Burnout state; motivation decline, weak self-regulation.

(a) Potential self-discipline learners

The pedagogical goal for this type of learner is to enhance the depth of participation and strengthen the practical application, so it is necessary to build a hierarchical task and collaborative environment to promote in-depth participation. In terms of task strategy, these learners have high learning ability and strong self-discipline but limited experience in collaboration, so advanced collaborative tasks should be designed, and complex tasks can stimulate the potential of teamwork. In terms of activity strategies, teacher-student interaction and peer interaction activities should be increased. For example, increase the number of open-ended questions to improve behavioral engagement and verbal expression; increase group collaborative activities, create online and offline interactive environments to provide students with opportunities to share and discuss, stimulate students' desire to express themselves and ask questions, and increase intrinsic motivation to learn. In terms of resource strategy, we should stimulate their potential as much as possible and push them to expand their knowledge and insights with extended resources. For example, industry white papers and videos of international academic conferences should be pushed to strengthen the connection between theory and reality. In terms of evaluation strategies, process evaluation should be increased, and the frequency of contributions to discussions should count toward the final grade in order to increase the motivation of participation.

(b) High expectations of low-effective learners

The teaching goal for this type of learner is to break the cycle of procrastination and consolidate the foundation of execution, and they need to rebuild their learning habits through structured tasks and instant feedback. In terms of task strategy, since this type of learner is weak in execution, it is necessary to break down the learning objectives and lower the threshold of action through fragmented tasks. In terms of activity strategies, progress monitoring and peer supervision groups are added to minimize delays, and individual coaching interventions are added. In terms of resource strategies, more resources can be used in the form of microclasses to increase students' interest in learning and level of understanding, while microclasses are short and concise to consolidate basic skills through a sense of instant achievement, supplemented by supervision and reminder strategies to increase engagement and concentration. In terms of evaluation strategies, process monitoring is strengthened.

(c) Active and participatory learners

The pedagogical goal for this type of learner is to transform ineffective engagement and build goal orientation, which requires activating deep learning through goal-oriented activities. In terms of task strategy, to address the problem of shallow engagement and low knowledge retention, deep engagement should be forced through structured paths. In terms of activity strategy, critical thinking is stimulated through competitive activities. In terms of resource strategy, these learners are behaviorally active but show no obvious preference in information processing and information input, and the recommendation of resources is based on video, adding practice exercises after the video to seize the excitement period to consolidate and strengthen the content. In terms of evaluation strategy, the number and content of effective contributions to discussions constitute the process grade to enhance the sense of achievement in learning.

(d) Negative burnout type learners

The instructional goals for this group of learners are to restart motivation and provide emotional support, and confidence needs to be rebuilt through resilience support and emotional intervention. In terms of task strategies, because of their weak resilience to stress, frustration needs to be reduced through progressive goals. In terms of activity strategy, task-completion goals should be set during lectures and scaffolding should be provided, such as task sheets, mind maps, learning clues and strategies, to help learners complete their tasks; learners should also discuss and exchange ideas with other students during learning activities and be required to speak, to increase behavioral interaction. In terms of resource strategies, core knowledge points are compressed into short

videos, and small successes are accumulated through low-difficulty tasks. In terms of evaluation strategies, learning motivation is gradually restored through positive incentives.

5. Qualitative Validation of Intervention Strategies via Expert Review

To address the need for empirical validation of the proposed intervention strategies prior to large-scale implementation, we conducted a qualitative validation study using the Delphi method. A panel of five experts was convened, consisting of three senior instructional designers from Hengshui Open University and two professors specializing in educational technology from the University of Science and Technology Beijing. The experts were tasked with evaluating the four proposed intervention strategies based on three criteria: (1) Pertinence (alignment with the identified persona characteristics); (2) Feasibility (ease of implementation within the Canvas LMS); and (3) Predicted Impact (potential to improve learner engagement).

The validation process involved two rounds of review. In the first round, experts provided open-ended feedback. For the 'Potential Self-Discipline' group, experts suggested that while collaborative tasks are beneficial, strict role definitions are necessary to prevent social loafing. For the 'Negative Burnout' group, the panel emphasized the importance of automated emotional support over purely instructional interventions. Based on this feedback, the strategies were refined.

In the second round, experts rated the finalized strategies on a 5-point Likert scale. The results indicated a high level of consensus regarding the efficacy of the framework. The 'Micro-task disassembly' strategy for the 'High Expectation Inefficiency' persona received the highest Feasibility score (4.8/5.0), confirming its practical value for addressing procrastination. The expert panel concluded that the multidimensional intervention framework offers a considerably more targeted approach than traditional "one-size-fits-all" methods, providing a robust theoretical basis for future quantitative experiments.

Table 15. Expert ratings of the four intervention strategies (second round, $n = 5$)

Learner persona	Intervention strategy component	Pertinence (1–5)	Feasibility (1–5)	Predicted Impact (1–5)	Inter-rater agreement (Kendall's W)
Potential self-discipline learners (Group 1)	Advanced collaborative tasks + expanded resources	4.6 (SD = 0.55)	4.2 (SD = 0.84)	4.4 (SD = 0.55)	0.78
High expectations of low-effective learners (Group 2)	Micro-task disassembly + progress monitoring	4.8 (SD = 0.45)	4.8 (SD = 0.45)	4.6 (SD = 0.55)	0.82
Active and participatory learners (Group 3)	Structured task chaining + competitive activities	4.4 (SD = 0.55)	4.0 (SD = 0.71)	4.2 (SD = 0.84)	0.74
Negative burnout type learners (Group 4)	Progressive goal setting + scaffolded guidance	4.6 (SD = 0.55)	4.4 (SD = 0.55)	4.4 (SD = 0.89)	0.76

6. Discussion

This study set out to develop a framework for precision teaching by modeling multidimensional learner personas. Our findings yield several key points for discussion regarding both the methodology of learner profiling and the nature of student engagement in online learning environments.

6.1. The superiority of Mahalanobis distance

The most significant methodological finding is the superior performance of the Mahalanobis distance over the Euclidean distance for clustering learner personas (Silhouette Score: 0.4277 vs. 0.3281). This result confirms our hypothesis that Mahalanobis distance, by design, is better suited to handle the correlated, multidimensional data characteristic of our persona model. While a direct comparison of Silhouette Scores across different studies and datasets can be misleading, a score of 0.4277 indicates a reasonable and well-defined cluster structure, strengthening the validity of the identified personas. This finding serves as an important empirical validation for educational data mining practitioners, emphasizing the benefits of selecting distance metrics that align with the underlying structure of the data.

6.2. Comparison with existing methods

Unlike studies that rely solely on density-based methods like DBSCAN, which we found to underperform on this dataset (Silhouette Score = 0.1532), our Mahalanobis-enhanced K-Means++ approach (Silhouette Score = 0.4277) effectively captures the globular but correlated nature of learner data. Furthermore, compared to the single-dimensional persona models criticized in the literature (*e.g.*, relying only on grades), our multidimensional approach aligns better with the complex reality of student engagement, as evidenced by the distinct separation of the 'High Expectation Inefficiency' group — a segment often overlooked by performance-only clustering.

To further verify the superiority of the proposed Mahalanobis-enhanced K-Means++ algorithm in educational data mining scenarios, we additionally compared it with two other representative clustering algorithms commonly used in learner profiling: Gaussian Mixture Models (GMM) and Spectral Clustering. The comparison was conducted using the same Canvas dataset and evaluation metrics, with hyperparameters optimized for each algorithm (*e.g.*, GMM used the Bayesian Information Criterion (BIC) to select the number of components, and Spectral Clustering used the normalized Laplacian matrix with $k = 4$ clusters).

1. The proposed algorithm outperforms GMM in all three metrics (Silhouette Score: 0.4277 vs. 0.3812; Davies-Bouldin Index: 0.8118 vs. 0.8543; Calinski-Harabasz Index: 1906.56 vs. 1623.49). GMM assumes that data follows a Gaussian distribution, which does not fully match the non-normal characteristics of educational behavioral data (*e.g.*, skewed distribution of interaction frequency), leading to suboptimal clustering results.
2. Spectral Clustering achieves a slightly lower Silhouette Score (0.3956) than the proposed algorithm and a significantly lower Calinski-Harabasz Index (1289.71). Although Spectral Clustering is effective for non-linear data structures, it has high computational complexity (especially for large datasets with 325,000 records) and is prone to overfitting when dealing with high-dimensional educational data, which limits its practical application in smart teaching platforms.

This comparison confirms that the Mahalanobis-enhanced K-Means++ algorithm balances clustering accuracy, computational efficiency, and adaptability to educational data characteristics, making it more suitable for large-scale learner persona modeling than other mainstream baseline models.

6.3. Interpretation of learner personas

The four personas identified — 'Potential Self-Discipline', 'High Expectation Inefficiency', 'Active Participation', and 'Negative Burnout' — provide a nuanced snapshot of the student population. These archetypes resonate with findings from higher education research

on student engagement and motivation. For example, the 'High Expectation Inefficiency' persona mirrors the well-documented psychological phenomenon of the intention-behavior gap, where students' stated goals and ambitions do not align with their actual learning actions. Similarly, the 'Negative Burnout' persona reflects patterns of declining motivation often observed in longitudinal studies of online learners. These personas move beyond simple performance-based groupings to capture the underlying drivers and barriers to learning, offering educators more precise targets for intervention.

6.4. Implications for precision teaching

The ultimate goal of this work is to enable precision teaching. The proposed intervention strategies, while preliminarily validated by expert review, are directly tied to the specific problems identified for each persona. For instance, the strategy for 'Active Participation' learners focuses on transforming their high but unfocused activity into deep, goal-oriented learning. This aligns with educational theories emphasizing the importance of metacognitive strategies and self-regulation. By diagnosing not just what a student is doing, but potentially why, educators can move from generic support to truly personalized guidance. However, it is crucial to note that the efficacy of these strategies in a real-world teaching environment has not yet been empirically tested, which constitutes a primary direction for our future work.

7. Limitations and Future Work

While this study provides a novel framework for multidimensional learner persona modeling, it is important to acknowledge its limitations, which in turn suggest avenues for future research.

The primary limitation is the lack of empirical validation for the proposed intervention strategies in an authentic classroom setting. As explicitly discussed in Section 5, the strategies have only undergone a preliminary Delphi expert review for qualitative feasibility assessment. A planned quasi-experimental study will be necessary to evaluate their actual impact on learning outcomes and engagement.

A second limitation pertains to the data source. The study utilized the Canvas open dataset, which primarily contains static behavioral and course completion data. The "Motivational-affective" dimension was operationalized using LMS-derived proxy variables (user type, stated objectives, expected hours), which may not be fully aligned with direct psychological or physiological measures of affect. This limitation serves as a clear boundary for our model's current interpretive scope.

To address these limitations, our future research agenda will proceed along two main tracks. First, we plan to conduct a quasi-experimental study to validate the proposed intervention strategies. This will involve implementing the persona-based interventions in a live course with an experimental group, while a control group receives standard instruction. Second, we aim to enhance the persona model by incorporating multi-modal and real-time data streams, such as trace-level log data and, where ethically and practically feasible, self-reported experience sampling methods, to move closer towards capturing authentic emotional and cognitive states in real time.

8. Conclusion

Taking learner persona as the core, this study constructs an analysis model integrating behavioral, cognitive and emotional dimensions, and achieves accurate clustering through the Mahalanobis-enhanced K-Means++. The experimental results show that the clustering algorithm based on Mahalanobis distance performs better in multi-indicator evaluation, which verifies its effectiveness in dealing with multi-dimensional correlated data. Through the empirical analysis of Canvas open dataset, the study classifies college learners into four typical groups: potential self-discipline type, high expectation low efficiency type, active participation type and negative burnout type, which reveals the core contradictions of each type of group in terms of goal orientation, behavioral commitment and motivational expectations state. The study proposes a hierarchical intervention strategy to address the characteristics of different groups:

the potential self-discipline group needs to strengthen practical ability through interdisciplinary collaborative tasks and expanding resources; the high-expectation and low-efficiency group needs to rebuild its execution ability through micro-task disassembly and progress monitoring; the active participant group needs to activate in-depth learning by using structured task chaining and competition; and the negative burnout group relies on resilient paths and emotional support to restore its motivation to learn. The proposed intervention strategies have undergone a preliminary qualitative validation through an expert Delphi review, confirming their pertinence and feasibility. Their application effectiveness in authentic teaching scenarios, however, requires further empirical investigation, which forms the basis for our next stage of research.

The contributions of this study are (1) breaking through the paradigm of single-dimension persona construction and proposing a dynamic adaptive multidimensional model to provide a more comprehensive analysis framework for accurate teaching; (2) solving the limitations of traditional K-means++ in data relevance modeling by improving clustering algorithms, which enhances the accuracy of subgrouping and intervention targeting; and (3) providing a practical intervention program for college teachers.

However, there are still some limitations in this study: first, the data source is limited to the static behavioral data of Canvas platform, and real-time affective computing can be integrated in the future to enhance the dynamics of the persona; second, the experimental data does not cover interdisciplinary and multi-segment scenarios, and the universality of the model needs to be verified in the future; finally, the evaluation of intervention strategies in this study remains at the expert review level, and long-term educational effectiveness in real classrooms still needs to be tracked and analyzed through future quasi-experimental studies. Future research can further explore the integration of learner profiling and educational knowledge mapping, and realize adaptive intervention by combining reinforcement learning technology.

Declaration of Competing Interests

The authors declare no conflict of interest

Funding

This research was funded by Teaching Talent Training Program of University of Science and Technology Beijing under grant no. JXGG202510.

Data availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

References

[1] R. Sahar *et al.*, "Artificial intelligence in sustainable education: A bibliometric analysis and future research directions," *Education Science and Management*, vol. 3, no. 1, pp. 57–77, 2025. https://doi.org/10.56578/esm030105

[2] R. Xia *et al.*, "The construction of student-centered artificial intelligence online music learning platform based on deep learning," *Scientific Reports*, vol. 15, no. 1, p. 15539, 2025. https://doi.org/10.1038/s41598-025-95729-w

[3] H. Xu *et al.*, "Research on constructing a smart classroom for computer general education courses centered on students," *Education Journal*, vol. 8, no. 1, pp. 50–57, 2025. https://doi.org/10.31058/j.edu.2025.81007

[4] X. Xiao and H. Lv, "Foundations, manifestations, and pathways of digital technology empowerment in higher education ideological and political theory courseware," *International Journal of New Developments in Education*, vol. 6, no. 12, pp. 133–141, 2024. https://doi.org/10.25236/IJNDE.2024.061220

[5] H. J. He *et al.*, "Fire risk assessment and simulation of multi-functional teaching buildings: A combined application of DEMATEL and PyroSim," *Journal of Engineering Management and Systems Engineering*, vol. 4, no. 1, pp. 1–20, 2025. https://doi.org/10.56578/jemse040101

[6] C. Griffin and L. Mannion, "Looking towards the future: The potential of precision teaching in the work of the educational psychologist in Ireland," *Tizard Learning Disability Review*, vol. 30, no. 1, pp. 26–30, 2025. https://doi.org/10.1108/TLDR-02-2024-0014

[7] K. Ferris and M. Fabrizio, "Foundations and frontiers in precision teaching: Purpose, standards, technology, and community," *Tizard Learning Disability Review*, vol. 30, no. 1, pp. 12–16, 2025. https://doi.org/10.1108/TLDR-10-2024-0037

[8] A. Ozturk and A. T. Kumtepe, "Relationship between learner profiles and learner–content interaction in online learning: Exploring implications for learning experience design," *Distance Education*, vol. 44, no. 3, pp. 425–457, 2023. https://doi.org/10.1080/01587919.2023.2226621

[9] T. Duan, "Categorization algorithm of vocal music online course resources under the background of MOOC teaching," *International Journal of High Speed Electronics and Systems*, vol. 254, p. 0484, 2025. https://doi.org/10.1142/S012915642540484X

[10] A. Cooper, *The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity*, Sams, 2004.

[11] D. Luis *et al.*, "Generic service integration in adaptive learning experiences using IMS learning design," *Computers & Education*, vol. 57, no. 1, pp. 1160–1170, 2011. https://doi.org/10.1016/j.compedu.2010.12.007

[12] J. Li, "2017–2024 China learner portrait research status and hot spot analysis: Citespace visualization study based on CNKI," *World Education Forum*, vol. 2, no. 4, pp. 52–59, 2024. https://doi.org/10.18686/wef.v2i4.4597

[13] W. Feng *et al.*, "Research on the construction and application of intelligent tutoring system for English teaching based on generative pre-training model," *Systems and Soft Computing*, vol. 7, p. 200232, 2025. https://doi.org/10.1016/j.sasc.2025.200232

[14] F. Alshaikh and N. Hewahi, "Intelligent tutoring system: A pedagogical model approach based on deep reinforcement learning," *IAENG International Journal of Computer Science*, vol. 52, no. 4, pp. 1196–1212, 2025.

[15] R. M. Suleman *et al.*, "A new perspective of negotiation-based dialog to enhance metacognitive skills in the context of open learner models," *International Journal of Artificial Intelligence in Education*, vol. 26, pp. 1069–1115, 2016. https://doi.org/10.1007/s40593-016-0118-8

[16] R. B. R. Maddu and S. Murugappan, "Online learners' engagement detection via facial emotion recognition in online learning context using hybrid classification model," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 43, 2024. https://doi.org/10.1007/s13278-023-01181-x

[17] L. Xu, "Research on visualization system of student behavior analysis based on campus big data," in *Proceedings of the 3rd International Conference on Public Management and Big Data Analysis (PMBDA 2023)*, Nanjing, China, 2023. https://doi.org/10.4108/eai.15-12-2023.2345331

[18] H. T. H. Duong *et al.*, "Academic performance warning system based on data driven for higher education," *Neural Computing and Applications*, vol. 35, no. 8, pp. 5819–5837, 2023. https://doi.org/10.1007/s00521-022-07997-6

[19] S. Zhang, M. Shen and Y. Yu, "Research on student portrait system based on canopy and k-means algorithm," in *Proceedings of the International Conference on Computer Application and Information Security*, vol. 12260, pp. 486–492, 2022. https://doi.org/10.1117/12.2637367

[20] C. Binder, "Precision teaching: Measuring and attaining exemplary academic achievement," *Youth Policy*, vol. 10, no. 7, pp. 12–15, 1988.

[21] H. Duan, "Research and practice of digital training model based on artificial intelligence technology," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, pp. 1–17, 2024. https://doi.org/10.2478/amns.2023.2.01507

[22] G. Siemens and R. S. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pp. 252–254, 2012. https://doi.org/10.1145/2330601.2330661

[23] J. He *et al.*, "A Data-Driven Innovation Model of Big Data Digital Learning and Its Empirical Study," *Information Dynamics and Applications*, vol. 1, no. 1, pp. 35–43, 2022. https://doi.org/10.56578/ida010105

[24] L. Mannion, "The use of Precision Teaching as an intervention for improving the skill acquisition of children with autism," *Educational Psychology in Practice*, vol. 38, no. 1, pp. 57–74, 2022. https://doi.org/10.1080/02667363.2022.2033956

[25] Y. Huang *et al.*, "A method for classroom behavior state recognition and teaching quality monitoring," *International Journal of Intelligent Computing and Cybernetics*, vol. 18, no. 2, pp. 382–396, 2025. https://doi.org/10.1108/IJICC-11-2024-0561

[26] M. Podoliak *et al.*, "Evaluating the impact of AI-based tools on language proficiency and motivation: Experimental evidence from philology students in Ukraine," *Journal of Research, Innovation and Technologies*, vol. 4, no. 3, pp. 271–282, 2025. https://doi.org/10.56578/jorit040303

[27] M. Al-Murtadha, "Motivating second language learners with a possible selves intervention program," *Journal of Language, Identity & Education*, vol. 24, no. 1, pp. 43–56, 2025. https://doi.org/10.1080/15348458.2022.2075874

[28] M. Gui *et al.*, "Mobile media education as a tool to reduce problematic smartphone use: Results of a randomised impact evaluation," *Computers & Education*, vol. 194, p. 104705, 2023. https://doi.org/10.1016/j.compedu.2022.104705

[29] W. Diao *et al.*, "Research on the construction of smart learning models supported by artificial intelligence," *International Journal of High Speed Electronics and Systems*, vol. 2025, p. 2540396, 2025. https://doi.org/10.1142/S0129156425403961

[30] B. Liu and D. Yuan, "Research on personalized teaching strategies based on learner profiles in a blended learning environment," *International Journal of Information and Communication Technology Education*, vol. 20, no. 1, pp. 1–23, 2024. https://doi.org/10.4018/IJICTE.346823

[31] L. Zheng *et al.*, "The effectiveness of the flipped classroom on students' learning achievement and learning motivation: A meta-analysis," *Educational Technology & Society*, vol. 23, no. 1, pp. 1–15, 2020.

[32] X. Zhang, "An innovative model of higher mathematics curriculum education incorporating artificial intelligence technology," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, pp. 1–16, 2024. https://doi.org/10.2478/AMNS.2023.2.01524

[33] M. Mujahid, "Mukhtar Yahya and the method of teaching Arabic: Analysis of Muhadasah method in the perspective of behaviorism theory," *Jurnal Pendidikan Islam*, vol. 7, no. 1, pp. 101–120, 2018. https://doi.org/10.14421/jpi.2018.71.103-122

[34] Z. Zhao and Z. Wei, "A study on behavioural characterisation of music learners and optimisation of teaching strategies based on data mining," *Applied Mathematics and Nonlinear Sciences*, vol. 10, no. 1, pp. 1–17, 2025. https://doi.org/10.2478/AMNS-2025-0019

[35] G. M. Haslerud, *Transfer, Memory, and Creativity: After-Learning as Perceptual Process*, University of Minnesota Press, 1972.

[36] M. R. Gharib *et al.*, "Examining the role of empowerment criteria on employee performance: A quantitative analysis in the oil industry," *Journal of Engineering Management and Systems Engineering*, vol. 2, no. 2, pp. 96–107, 2023. https://doi.org/10.56578/jemse020201

[37] Z. Li *et al.*, "Knowledge Interaction Failure in Virtual Communities: Modeling the Impact of Value Co-Destruction on Knowledge System Performance," *International Journal of Knowledge and Innovation Studies*, vol. 3, no. 3, pp. 129–141, 2025. https://doi.org/10.56578/ijkis030301

Contact addresses:

Yafeng Liu
Further Education Department
Hengshui Open University
Hengshui, China
e-mail: liuyf20200225@163.com

Rui Yan
University of Science and Technology Beijing
Beijing, China
e-mail: yanrui@ustb.edu.cn

Yong Xiao
University of Science and Technology Beijing
Beijing, China
e-mail: 2672107562@qq.com

Ailing Chen
University of Science and Technology Beijing
Beijing, China
e-mail: x20051128y@163.com

Di Zhang*
Further Education Department
Hengshui Open University
Hengshui, China
e-mail: hskfdxzd112233@163.com
*Corresponding author

Yafeng Liu received her bachelor's degree from Hengshui Open University, where she is currently a lecturer. She has led or participated in several municipal-level research projects and has published seven papers, one of which is indexed in the Web of Science. Her work covers areas such as art education, higher education, and low-carbon development.

Rui Yan received his Ph.D. in Management Science and Engineering from the National University of Singapore. He is currently a professor and master's supervisor at the School of Economics and Management, University of Science and Technology Beijing, China. His main research interests include mission reliability, energy system reliability, and optimization modeling.

Yong Xiao is an undergraduate student at the University of Science and Technology Beijing, majoring in Information Management and Information Systems. His main research interests are operations research, optimization, and operations management.

Ailing Chen is a graduate student at the University of Science and Technology Beijing, majoring in Technology Economics and Management. Her main research interest is the operation and management of the low-carbon economy.

Di Zhang received her bachelor's degree in 2006. She worked at Hengshui Fifth People's Hospital from 2009 to 2017 and has worked at Hengshui Open University since 2017. She has led or participated in several provincial and municipal research projects and has published ten papers, two of them in Chinese core journals. Her work covers topics such as nursing, mental health, and higher education.

Received: March 2026
Revised: April 2026
Accepted: May 2026