International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 17 (2025) pp. 1-14

https://doi.org/10.70917/ijcisim-2025-0183
Article

Research on Quality Monitoring and Data Mining Methods for Graduate Education in Innovation Management

Man Luo *

South China Normal University, Guangzhou, Guangdong 510631, China; hslm629@163.com

Abstract: With the continuous growth in the scale of postgraduate admissions, monitoring the quality of postgraduate education has become an urgent issue that needs to be addressed. To this end, this paper establishes an information platform for monitoring the quality of postgraduate education from three aspects: structural design, functional design, and permission design. Based on the data analysis of this platform, data mining methods are used to study the learning behaviour and learning effectiveness of postgraduates. Using an improved K-means algorithm, graduate students are categorised into four types, with knowledge-exploration-oriented and marginally passive-oriented types being the most common, accounting for 37.99% and 27.60%, respectively. Multivariate regression analysis is then used to construct a multiple linear regression equation linking graduate students' learning outcomes to their learning behaviour characteristics. The regression coefficients for the number of knowledge tests and the module completion rate are 0.259 and 0.217, respectively, making them the primary factors influencing graduate students' academic performance. Utilising big data to analyse and predict graduate students' learning behaviours and learning outcomes facilitates monitoring the quality of graduate education and comprehensively improving the management level of graduate education.

Keywords: K-means; multiple regression analysis; learning behaviour analysis; data mining; graduate education quality

1. Introduction

The cultivation of innovative capabilities among postgraduate students is the core component of postgraduate education in higher education institutions and an integral part of the national strategies for science and education-driven development and talent-driven national strength. Postgraduate students in the new era, who are at the most creative stage of their lives, should proactively assume the mission of innovative development, fully utilise the learning opportunities available during their time at university, and cultivate and shape their innovative capabilities and competencies [1-3]. Western countries such as the United Kingdom, the United States, and Canada have continuously implemented educational innovations since the end of the Second World War, actively promoting reforms in educational models and teaching methods for cultivating innovative talent, and continuously enriching theinner meaning of graduate education in fostering innovative capabilities [4-6]. The United States has established a relatively well-developed training mechanism for graduate education, emphasising the role of mentors in the training process and focusing on the cultivation of capabilities and innovative spirit. In Germany, graduate education adheres to strict standards for selecting mentors, emphasises the integration of industry, academia, and research, and encourages innovation. Japan implements a collective training system centred on professors, guided by the concept of innovation-driven science and technology. In China, most graduate education follows a demand-oriented, theory-practice integrated innovation training system under the guidance of mentors [7-9].

Some universities have a high volume of innovation project applications and innovation-related paper publications, but the actual innovation outcomes conversion rate is less than 50%, with some even at 0%.

This phenomenon highlights the shortcomings in the current graduate innovation quality management system, where most innovation evaluations rely solely on the number of published papers, lacking assessments from multiple perspectives such as innovative thinking and cross-disciplinary innovation. Additionally, the neglect of data usage and the absence of relevant innovation quality early warning mechanisms have made it difficult for innovation projects to progress, with the quality conversion of innovation project outcomes taking a long time, and quality feedback lagging behind changes in innovation demands [10-14]. de Souza Fleith and Gomes [15] introduced the effectiveness of the 'Creativity Teaching Practice Scale in Higher Education' in assessing graduate students' creativity, which can effectively evaluate students' general thinking, innovative thinking, and interest-based thinking. However, the scale assessment is subject to subjective influences. Pengjun [16] developed a model for evaluating graduate students' research and innovation capabilities based on a computer-simulated distributed identity management system, enhancing the efficiency and objectivity of the assessment. However, this method faces challenges related to technical integration and data security.

With the large-scale development of higher education, postgraduate education has also entered a phase of rapid expansion. How to enhance the innovative quality of students and ensure and continuously improve the quality of postgraduate education has become a hot topic. The establishment and improvement of a quality monitoring system for postgraduate education have become the primary issue facing the management of high-level talent innovation. Logachev et al. [17] proposed an information system for monitoring and managing the quality of educational programmes and an objective quality assessment system, which can efficiently supervise the entire life cycle of educational programmes. Oseredchuk et al. [18] pointed out that the current quality monitoring of higher education has three levels of problems: thesis work decisions, monographs and textbooks, and scientific journals. These problems reflect the current shortcomings in theses and academic materials, which are important supports for student innovation management.

Li and Zhang [19] utilised big data technology to monitor the quality of higher education teaching and analysed teaching quality monitoring and related assessment factors through the K-means clustering algorithm. They introduced an association rule mining algorithm to analyse the clustering results, mine the correlations between data, and construct a teaching quality monitoring and assessment model, significantly improving the accuracy and effectiveness of teaching quality monitoring. Education has now entered the digital transformation phase, with data-driven management models becoming a focal point. Various digital technologies are being used to achieve educational quality assessment, educational quality monitoring, and the construction of student data platforms, with data mining technology playing a crucial role in this process [20-22]. Wang [23] combined 6G internet communication technology with data mining technology to optimise teaching quality monitoring and assessment. This method can precisely evaluate students' learning status and teaching quality, thereby managing classroom resources.

The study, based on quality monitoring dimensions, proposes an information platform for monitoring the quality of postgraduate education, explaining it from three aspects: structural design, functional design, and permission design. Subsequently, for the purpose of learning behaviour analysis at the platform's business layer, online learning characteristic data of postgraduate students from a certain university were collected, and Pearson correlation analysis was used to screen learning characteristics. After improving the differential evolution algorithm by introducing adaptive operators, a multi-variation strategy with weight coefficients, and Gaussian perturbation crossover operations, the improved differential evolution algorithm was used to optimise the K-Means clustering algorithm, which was then applied to postgraduate learning behaviour analysis to obtain the learning behaviour profiles of the selected postgraduate students. Using multiple linear regression methods, the relationship between graduate students' learning behaviour characteristics and academic performance was explored, and a regression equation was constructed. Finally, the effectiveness of the constructed learning performance prediction model was tested using histograms of standardised residuals and normal P-P plots.

2. Information platform for monitoring the quality of postgraduate education

The widespread use of big data-driven technologies has promoted epistemological and methodological changes in various fields such as society, economy, education, culture, and science and technology, triggering changes in educational management methods. Quality monitoring of postgraduate education is a product of educational management development in the big data era. Therefore, an information platform for quality monitoring of postgraduate education should be constructed based on innovative management.

2.1. Structural Design

The Graduate Education Quality Monitoring Information Platform is a cross-platform, web-based

application system that leverages the high concurrency, high speed, multi-dimensional, and cross-platform characteristics of computer technology and information-based web platforms. By collecting, tracking, recording, and electronically archiving graduate student training data from multiple dimensions, the system constructs a data model for the graduate education quality monitoring system, performs statistical analysis, classification, and correlation analysis on the data, and establishes a stable, data-closed, and digitally-enabled graduate education quality monitoring system. It also achieves data integration and sharing for graduate education resource management, as well as unified management and coordination of information resources. The system adopts an MVC architecture for layered implementation, with the system interface layer utilizing web front-end technology to present platform content. The business layer uses the SSM framework to implement the processing workflows of sub-systems. Data interaction is based on the MySQL database. Nginx is used to separate the frontend and backend, ensuring load balancing and reducing coupling. The architectural diagram of the graduate education quality monitoring information platform is shown in Figure 1.

The architecture of the postgraduate education information platform is divided into three layers: the data layer, the business layer, and the view layer. The data layer mainly performs basic data operations and interfaces with other platforms to achieve data sharing. The business layer primarily handles specific business operations and data analysis tasks, such as student management, grade management, and degree management, which are considered basic business functions. Postgraduate learning behaviour, educational progress, and thesis quality fall under data analysis business functions. The view layer presents information to postgraduates or administrative departments, responds to operations, and pushes relevant information.

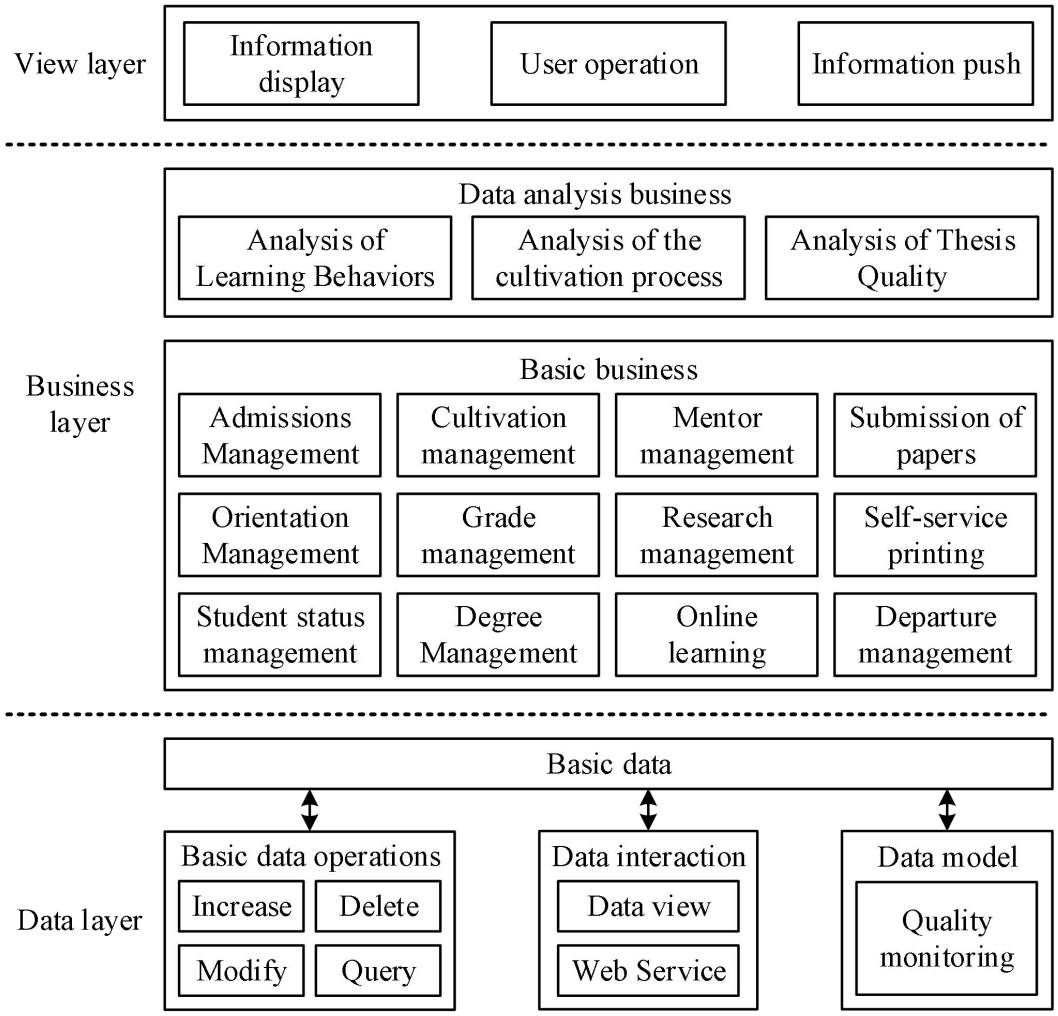

Figure 1. The architecture of the education quality monitoring information platform.

2.2. Functional Design

The Information Platform for Monitoring the Quality of Postgraduate Education is a vertical management structure coordinated by the Graduate School and managed by secondary colleges. It mainly consists of multiple business systems, including postgraduate enrolment management, education management, thesis review, orientation, and graduation systems. At the same time, to ensure data standardisation, it adopts the data interface standards specified by the Ministry of Education to facilitate data transfer and sharing between systems.

The Graduate Education Quality Monitoring Information Platform primarily consists of modules such as admissions management, orientation, educational management, course centre, thesis review submission, graduate self-service printing platform, and graduation system, covering all stages of a graduate student's journey from enrollment to graduation. The Graduate Education Quality Monitoring Management System serves as the core component of the entire information platform, featuring functions such as academic record management, programme management, grade management, degree management, advisor management, and graduate affairs management. The Course Centre provides graduate students with an online learning platform, enabling them to study without being constrained by their learning environment. The thesis review system enables online blind review of graduate theses. The self-service printing platform allows graduate students to print various certificates such as academic transcripts, enrollment certificates, and degree certificates at any time. Additionally, each module of the platform includes data import and export functions, ensuring that graduate student basic data can be imported according to templates and statistical information required for reporting can be quickly exported.

2.3. Permission Design

Graduate students, supervisors, and graduate education administrators are all users of the graduate education information platform. They log in through the university's unified service portal and are assigned different permissions based on their account roles to ensure information security. For example, super system administrators, as the overall system maintenance managers, are granted super permissions. The Graduate School, as the primary administrative department, is granted first-level permissions. Secondary colleges, as secondary management units, are granted secondary management permissions for their respective units and can only view and manage data related to the quality of graduate education within their own college. Other relevant functional departments are granted corresponding management permissions based on the scope of their responsibilities.

3. Learning Behaviour Analysis Based on Data Mining

To achieve learning behaviour analysis at the business layer of the informationised platform for monitoring the quality of postgraduate education, and to explore the correlation between postgraduate learning behaviour and learning outcomes, this chapter primarily uses an improved K-means clustering algorithm to analyse postgraduate learning behaviour.

3.1. Clustering Algorithms

To address the issues in the traditional K-Means clustering algorithm and improve clustering effectiveness and efficiency, an improved differential evolution algorithm was used to optimise the K-Means clustering algorithm and applied to the analysis of postgraduate learning behaviour.

3.1.1. Improved Differential Evolution Algorithm

(1) Adaptive operation operator

The mutation and crossover operations in the differential evolution algorithm are the core components of the entire algorithm, and the mutation factor F and crossover factor C_R have a significant impact on the algorithm. In traditional differential evolution algorithms, mutation factor F and crossover factor C_R are fixed values, which limits the algorithm's optimisation capability and convergence performance, hindering improvements in algorithm optimisation performance. During the early stages of algorithm evolution, it is important to maintain the diversity of population individuals to enhance the algorithm's global optimisation capability. In the later stages of the algorithm, it is necessary to strengthen the algorithm's local search capability to improve convergence speed.

In order to enable F and C_R to adapt to the optimisation requirements of the algorithm at different stages of evolution, this paper uses adaptive operation operators for improvement, which adaptively balance the global search capability and local search capability of the algorithm as it evolves. The

adaptive strategy of the operation operators is as follows:

$$F = F_{\max} - (F_{\max} - F_{\min})(G / G_{\max})^2 \tag{1}$$

$$C_R = C_{R\max} - (C_{R\max} - C_{R\min})(G / G_{\max})^2 \tag{2}$$

In the formula: $F_{\min}$ represents the lower limit of F, $F_{\min} = 0.3$. $F_{\max}$ represents the upper limit of F, $F_{\max} = 0.9$. $C_{R\min}$ represents the lower limit of C_R, $C_{R\min} = 0.3$. $C_{R\max}$ represents the upper limit of C_R, $C_{R\max} = 0.9$. G represents the current iteration count of the algorithm. $G_{\max}$ represents the maximum iteration count specified by the algorithm.

By implementing adaptive adjustments to F and C_R according to Equations (1) and (2), it is possible to ensure that mutation factor F and crossover factor C_R decrease linearly as the number of algorithm iterations increases.

(2) Multi-variant strategy introducing weight coefficients

Based on the characteristics of different mutation strategies, we combine the DE/rand/1 mutation strategy with the DE/current-to-best/2 mutation strategy to propose a differential evolution algorithm improved by multiple mutation strategies. By introducing a weight coefficient W, we achieve adaptive adjustment of the proportions of different mutation strategies according to the number of evolutionary iterations of the algorithm, thereby leveraging the advantages of different mutation strategies at different evolutionary stages of the algorithm.

$$W = W_{\min} + (W_{\max} - W_{\min})(G / G_{\max}) \tag{3}$$

$$\begin{aligned} V_{i,G+1} = {} & (1-W)[X_{r_1,G} + F(X_{r_2,G} - X_{r_3,G})] \\ & + W[X_{i,G} + F(X_{best,G} - X_{i,G}) \\ & + F(X_{r_4,G} - X_{r_5,G}) + F(X_{r_6,G} - X_{r_7,G})] \end{aligned} \tag{4}$$

In the formula: W is the weighting factor, $W_{\min} = 0, W_{\max} = 1$. G is the current iteration count. $G_{\max}$ is the maximum iteration count set by the algorithm. $V_{i,G+1}$ represents the mutated individuals of the $G+1$ generation. $X_{best,G}$ represents the individual with the best fitness value in the current population. $r_1, r_2, r_3, r_4, r_5, r_6, r_7$ represents seven random numbers taken from $[1, NP]$ that are not equal to i and are mutually distinct.

(3) Gaussian perturbation cross operation

To maintain diversity in the "dimension" of the population and avoid the algorithm falling into a local optimum, this paper introduces a Gaussian perturbation mechanism based on the best individual in the current population into the crossover operation of the differential evolution algorithm, using the best individual in the current population to guide the evolutionary direction of the other individuals. At the same time, Gaussian perturbation is used to randomly generate new values in each dimension with a certain probability, maintaining the diversity of the population in the "dimension." The specific operation is as follows.

1) For clustering problems, for a certain dimension n, take a random number $rand(0,1)$. If $rand(0,1) \le C_R$, then $U_{i,G}^n = V_{i,G}^n$, and then end the intersection operation for that dimension. Otherwise, proceed to step 2).

2) Take another random number $rand(0,1)$. If it is $rand(0,1) \le 0.7$, then it is $U_{i,G}^n = X_{best,G}^n$. If it is $rand(0,1) > 0.7$, then generate a random value in that dimension using Gaussian perturbation. The generation formula is:

$$U_{i,G}^n = X_{best,G}^n \cdot [1 + C \cdot N(0,1)] \tag{5}$$

In the equation: C is the control parameter for Gaussian disturbance, $C = 0.1$. Random value $N(0,1)$ represents a normal distribution with a mean of 0 and a standard deviation of 1.

3.1.2. Improved K-Means Algorithm

The improved differential evolution algorithm is integrated with the K-Means clustering algorithm. The fitness function value of each individual is calculated using the fitness function in equation (6). Iterative optimization is performed using the improved differential evolution algorithm. The optimal individual output at the end of the algorithm replaces the initial cluster centers randomly selected by the traditional K-Means clustering algorithm.

$$fitness = SSE = \sum_{ik} (x_{i,k} - c_k)^2 \tag{6}$$

In the formula: $x_{i,k}$ represents a data point, and c_k represents the center of the cluster to which the data point belongs. The smaller the SSE, the better the clustering effect. Conversely, the larger the SSE, the worse the clustering effect.

In clustering algorithms, the ultimate goal is to minimize the density of data points within a class and maximize the density between classes. Therefore, the sum of squared errors (SSE) is used as the fitness function for improving the differential evolution algorithm, as shown in Formula (6).

3.1.3. Algorithm Flow

(1) Set the population size NP to 10 times the solution dimension D, and adaptively determine the mutation factor F and crossover factor C_R according to Equations (1) and (2), $F_{\min} = 0.3$, $F_{\max} = 0.9$, $C_{R\min} = 0.3$, $C_{R\max} = 0.9$, with the current evolution number being G, the maximum evolution number of the algorithm being $G_{\max} = 1200$, and the Gaussian perturbation coefficient being $C = 0.1$.

(2) Initialize a population of size NP.

(3) For individual $X_{i,G}$ in the population, according to the multi-variation strategy with the introduction of weight coefficients, the variation individual $V_{i,G}$ is obtained.

(4) For parent individual $X_{i,G}$ and mutant individual $V_{i,G}$, intermediate experimental individual $U_{i,G}$ is obtained based on the Gaussian perturbation crossover operation.

(5) Calculate the fitness values of parent individual $X_{i,G}$ and intermediate experimental individual $U_{i,G}$ separately, and compare the fitness values. The individual with the higher fitness value will enter the next generation population to continue participating in evolution. Repeat steps 3) to 5) until the maximum number of evolutionary iterations specified by the algorithm is reached.

(6) Use the optimal individuals output by the algorithm as the initial cluster centers for K-Means clustering. Iterate the process until the clustering termination criteria are met or the maximum number of iterations is reached. The algorithm ends, and the clustering results are output.

3.2. Data Samples

This study took graduate students from a certain university as the research subjects. From the backend records of online courses, we randomly collected learning behavior data from 300 students during the spring semester of 2025. The data included various types of process-oriented and outcome-oriented information, such as basic information, online course learning status, interactive posts, practical training, test frequency, and assessment scores. After excluding missing samples, we obtained 279 valid samples. To ensure consistency in data units, the data was standardized using Z-score normalization. This method transforms the data into a standard normal distribution, where the mean is 0 and the standard deviation is 1.

3.3. Feature extraction

In the backend data of online courses, there are numerous metrics that describe student learning, including: Learning Performance X1, platform learning duration X2, online practice question-solving duration X3, micro-lecture viewing duration X4, chapter learning frequency X5, assignment submission frequency X6, discussion post frequency X7, platform practical exercise duration X8, active days X9, average assignment score X10, online test frequency X11, and module completion rate X12, among over 12 metrics.

The results of the correlation analysis for each indicator are shown in Figure 2, where * denotes p <

0.05 and ** denotes $p < 0.01$. To determine which indicators have greater guidance significance in the teaching process, Pearson correlation analysis was used to determine that the following six statistical indicators- "number of assignments submitted × 6," "number of discussion posts × 7", "Platform practical operation duration X8", "Average assignment score X10", "Number of knowledge tests X11", and "Module completion rate X12". These six statistical indicators were found to have a significant positive correlation $(p < 0.05)$ with students' final learning outcomes, making them statistically meaningful.

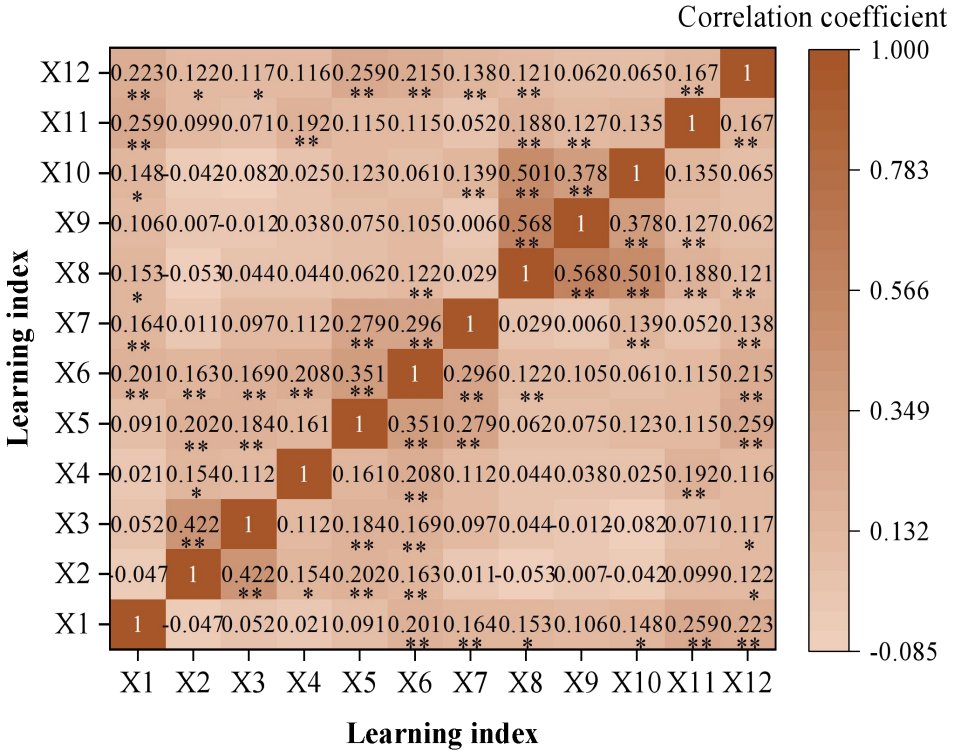

Figure 2. Correlation analysis results of each index.

3.4. K-means clustering analysis

3.4.1. Selection of the number of clusters

Based on the total number of samples, the range of K values was set to $[2,9]$. The contour coefficients for different K values are shown in Figure 3. When $K = 4$, the coefficient is at its maximum of 0.355, indicating the best clustering effect. $K = 5$ and $K = 3$ have the next highest coefficients after $K = 4$. Therefore, $K = 4$ was selected as the number of clusters for graduate student learning behavior.

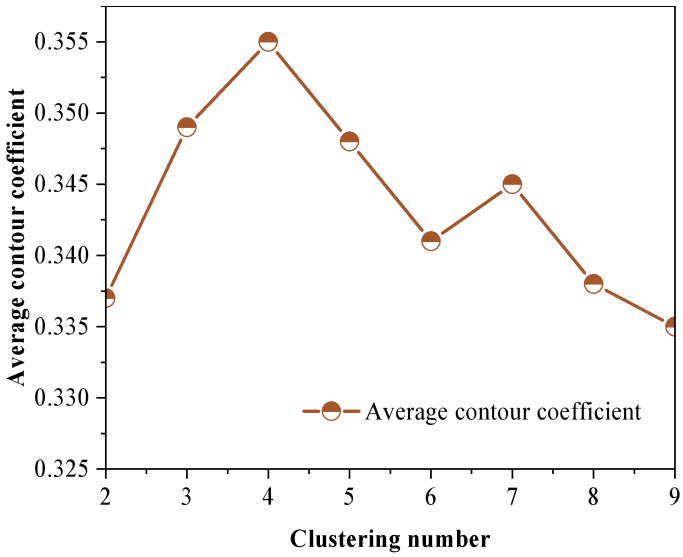

Figure 3. The contour coefficient of the different K values.

3.4.2. Clustering Results Analysis

The results of the variance analysis comparison of cluster categories are shown in Table 1. "Online test frequency X11" has the greatest importance for clustering, with a cluster contribution of 0.998, while "Discussion post frequency X7" has the least importance, with a cluster contribution of 0.174. Factors with importance exceeding 0.5 include "Online test frequency X11," "Assignment average score X10," "Number of Assignment Submissions X6", and "Platform Practical Training Duration X8". After obtaining the cluster categories, to explore the specific characteristics of each category, ANOVA was used to study the differences among the various category groups. The characteristics of graduate students in different categories across the six learning behavior research items all exhibited significant differences $(p < 0.01)$.

Table 1. Comparison results of variance analysis difference of cluster variance.

Feature term	Cluster contribution	Mean±SD				*F*	*P*
		A(n=35)	B(n=77)	C(n=61)	D(n=106)		
X11	0.998	17.58±2.49	9.76±2.08	9.14±2.45	9.71±1.58	43.172	0.007
X10	0.925	75.41±50.63	20.35±25.78	106.52±21.93	24.01±30.81	35.791	0.002
X6	0.722	15.93±3.64	7.95±3.94	12.17±4.22	14.49±2.51	64.285	0.005
X8	0.592	84.62±40.34	29.61±32.78	104.14±30.65	45.52±40.19	129.604	0.004
X12	0.258	16.09±5.69	8.37±6.53	10.58±6.93	15.17±6.29	41.427	0.001
X7	0.174	106.16±44.24	59.53±35.77	95.67±43.79	95.04±45.43	21.528	0.003

3.4.3. Student Profiles

Based on the clustering results, profiles were constructed for the four categories of students. The learner behavior profiles are shown in Figure 4. The four categories of graduate students, A, B, C, and D, account for 12.54%, 27.60%, 21.86%, and 37.99%, respectively. Group A comprises 35 students who performed exceptionally well in online tests, module completion, and discussion postings. These students have a solid foundation in professional knowledge, so Group A students are labeled as "professionally focused." Group B students had the lowest average scores in significant feature items, indicating that these students have weak self-management skills and spend relatively little time on learning, so they are labeled as "passive and marginalized." Group C comprises 61 students who spent significantly more time on assignment averages and platform practical tasks than students in other categories. They tend to focus on practical training and hands-on content, completing practical tasks quickly, and are labeled as "Practical Skills-Oriented." Based on Group D students' active performance in assignment submissions, module completion, and discussion post characteristics, these students are skilled at thinking and problem-solving and are labeled as "Knowledge Exploration-Oriented."

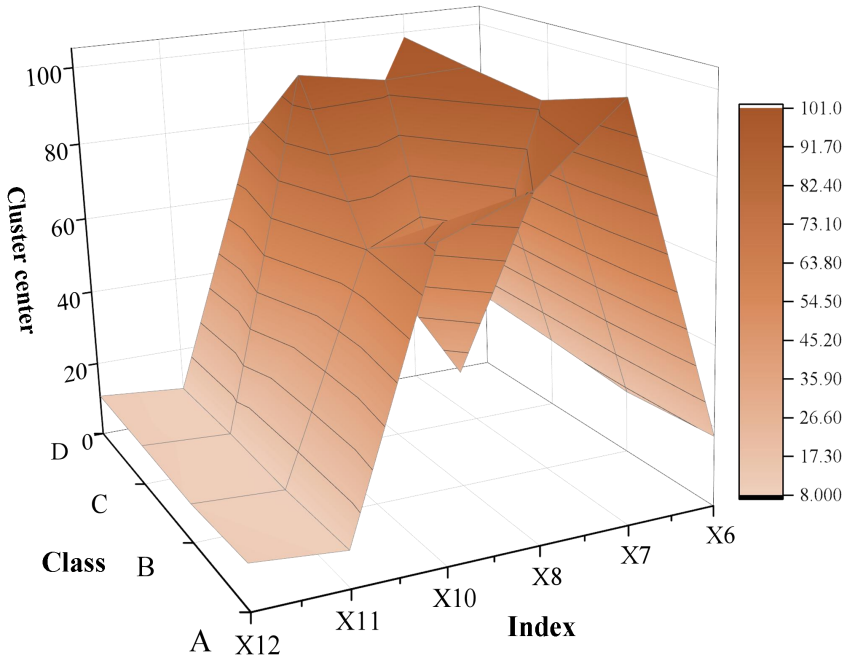

Figure 4. Learner behavior portrait.

4. Learning achievement prediction based on data mining

Based on the analysis of graduate students' learning behavior data to construct learner profiles, a learning achievement prediction model was developed using multiple linear regression methods to explore the relationship between graduate students' learning behavior and academic performance.

4.1. Multiple regression methods

Regression analysis is an important branch of modern applied statistics and a scientific method for studying the quantitative laws governing relationships between phenomena. It involves analyzing the interdependent relationship between a dependent variable and one or more explanatory variables, estimating or predicting the influence of explanatory variables on the dependent variable, and is a multivariate statistical analysis method for studying non-determinate relationships between variables. Regression analysis not only enables the analysis of the magnitude of the influence of explanatory variables on the dependent variable but also allows for the prediction and control of the dependent variable through regression equations.

4.1.1. Multiple linear regression model

When a dependent variable contains two or more independent variables, the model for multiple linear regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_m x_m + \delta \tag{7}$$

In the equation, $x_1, \cdots x_m$ is a non-random variable, β_0 is a constant term, $\beta_1, \beta_2 \cdots \beta_m$ is a regression coefficient, and δ is a random error term with a mathematical expectation equal to zero.

If n measurements are taken for y and x, n sets of observations $y_1, x_{1i}, \cdots, x_{mi} (i = 1, 2, \cdots, n)$ are obtained.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \cdots + \beta_m x_{mi} + \delta_i \tag{8}$$

Expressed as a matrix:

$$y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, x=\begin{bmatrix} 1 & x_{11} & \cdots & x_{m1} \\ 1 & x_{12} & \cdots & x_{m2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{mn} \end{bmatrix}, \beta=\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}, \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{9}$$

At this point, the model can be expressed as:

$$y = X\beta + \delta \tag{10}$$

δ is the error revealed between the data fitted by the model and the actual data.

4.1.2. Least squares method for finding the optimal solution

Take quadratic equations as an example. Given a set of sample data $x_i, y_i (i = 1,2,\cdots n)$, the regression function is required to fit this set of values as closely as possible. The criterion for finding the optimal solution using the least squares method is to minimize the sum of squared residuals.

The formula for the sum of squared residuals is:

$$Q = \sum_{1}^{n} (y_i - \hat{y}_i)^2 = \sum_{1}^{n} (y_i - ((\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \tag{11}$$

Take the partial derivatives of $\hat{\beta}_0$ and $\hat{\beta}_1$ separately. When the derivative is zero, Q takes its minimum value:

$$\frac{\partial Q}{\partial \beta_0} = 2\sum_{1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{12}$$

$$\frac{\partial Q}{\widehat{\partial \beta_1}} = 2\sum_{1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \tag{13}$$

Assuming there are more model variables $x_1, \cdots x_m$, the linear equation system can be represented by a matrix as follows:

$$\begin{bmatrix} 1 & x_{11} & \cdots & x_{m1} \\ 1 & x_{12} & \cdots & x_{m2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{mn} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{14}$$

The final optimal solution is:

$$\beta = (A^T A)^{-1} A^T Y \tag{15}$$

4.1.3. Model Parameters and Accuracy Verification

The mean square error (MSE) is generally used to test the accuracy of the optimal solution of a model. The mean square error refers to the expected value of the square of the difference between the estimated parameter value and the true parameter value, which is used to evaluate the degree of variation in the data. The smaller the MSE value, the higher the accuracy of the optimal solution of the prediction model. The calculation formula is:

$$MSE = \frac{SSE}{n} = \frac{1}{n}\sum_{i=1}^{m} w_i (y_i - \hat{y}_i)^2 \tag{16}$$

4.2. Learning Effectiveness Prediction Analysis

4.2.1. Model Validation

This section conducts a multiple linear regression analysis using six features as independent variables and graduate student academic performance as the dependent variable. Table 2 presents the test statistics used in the evaluation model, from which R, R^2, adjusted R^2, standard error of the estimate, and D-W statistic can be observed. In this study, the adjusted R^2 value of the regression model is 0.721, indicating that the model has a high degree of fit. Additionally, the D-W statistic is 1.699, indicating that the model's

residuals exhibit positive autocorrelation.

The results of the analysis of variance show that the F value of the regression part is 93.173, with a corresponding P value of 0.000, which is less than the significance level of 0.01. Therefore, it can be concluded that the six characteristics have a significant explanatory power for graduate student performance.

Table 2. Evaluation model inspection statistics.

R	R^2	Adjusted R^2	Standard estimation error	D-W statistics
0.854	0.735	0.721	8.378	1.699

4.2.2. Regression results

Table 3 shows the regression coefficients and other relevant statistics of the linear regression model. As can be seen from the table, the intercept term of the linear regression model is -7.721, and the significance levels of "number of assignments submitted X6," "number of discussion posts X7," "platform practical operation time X8," "average assignment score X10," "number of knowledge tests X11," and "module completion rate X12" are all less than 0.05, indicating that their coefficients are statistically significant. Therefore, the following regression equation can be constructed:

Academic performance=0.198*X6+0.165*X7+0.144*X8+0.138*X10 +0.259*X11 +0.217*X12-7.721.

Table 3. Regression coefficient of linear regression model and its relative statistics.

	Unnormalized coefficient		Normalized coefficient		Significance
	B	SE	*Beta*	*t*	
(intercept)	-7.721	5.721		-1.421	0.241
X6	0.198	0.008	0.254	0.844	0.008
X7	0.165	0.265	0.215	3.509	0.027
X8	0.144	0.044	0.193	6.411	0.035
X10	0.138	0.138	0.188	3.717	0.020
X11	0.259	0.231	0.305	5.445	0.012
X12	0.217	0.082	0.271	7.234	0.004

Table 4 lists several important residual statistics. From this table, key statistical information such as the minimum and maximum values of the predicted values, residuals, standardised predicted values, and standardised residuals can be extracted. Specifically, the maximum residual value is 14.849, the minimum is -44.392, and the average is 0.002. Figure 5 displays the histogram of standardised residuals, illustrating the frequency distribution of standardised residuals and clearly indicating that the standardised residuals generally follow a normal distribution. Figure 6 presents the normal P-P plot of standardised residuals. The horizontal axis of this plot represents the cumulative probability of the actual observed values, while the vertical axis represents the cumulative probability of the theoretical expected values. If the sample data follows a normal distribution, all data points should cluster around the diagonal line. Based on the results shown, it can be confirmed that the distribution indeed follows this pattern, further indicating that the standardised residuals approximately follow a normal distribution, consistent with the results shown in the histogram.

Table 4. Statistics of residues.

	Min	Max	Mean	Standard deviation
Predictive value	-5.921	98.569	84.262	13.972
Residual error	-44.392	14.849	0.002	8.246
Standard forecast	-8.618	0.926	0.000	1.000
Standard residue	-5.636	1.757	0.000	0.988

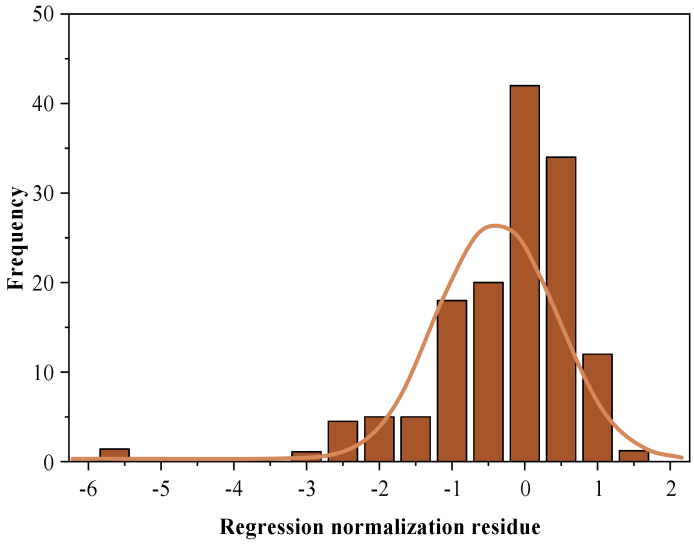

Figure 5. Standardized residual histogram.

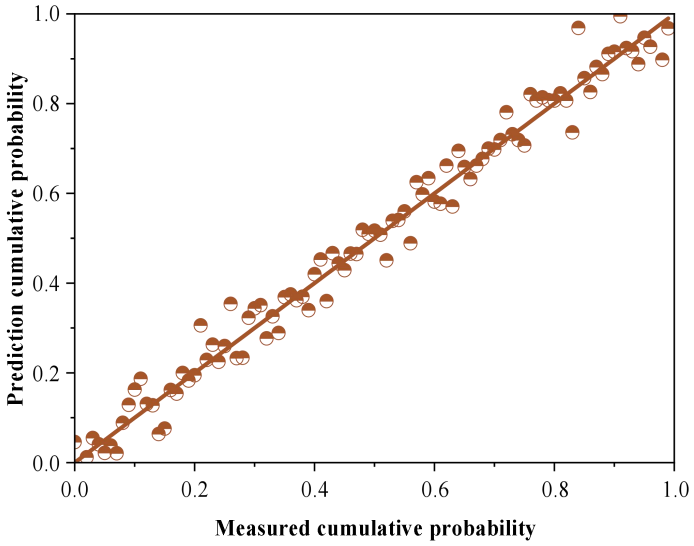

Figure 6. Standardized residual normal p-p diagram.

5. Conclusion

Innovating graduate education management methods based on data mining techniques holds significant importance for promoting the high-quality development of graduate education. This study focuses on innovative management, proposing an informationised platform for monitoring graduate education quality, and analysing graduate students' learning behaviour data. It employs improved K-means clustering and multiple regression analysis methods to explore graduate students' learning behaviour and predict learning outcomes.

By clustering different learning behaviours, postgraduate students were categorised into four types: professionally focused, marginally passive, practical skills-oriented, and knowledge-exploration-oriented. Among these, the largest proportions were found in the knowledge-exploration-oriented and marginally passive groups, accounting for 37.99% and 27.60% respectively. Additionally, a multiple linear regression equation was constructed to analyse the relationship between learners' learning outcomes and their learning behaviour characteristics. The results indicate that students' academic performance is primarily influenced by the number of knowledge tests taken, the proportion of modules completed, and the number of assignments submitted, with regression coefficients of 0.259, 0.217, and 0.198, respectively.

This project utilises data mining technology and statistical analysis methods to conduct an in-depth

analysis of postgraduate students' learning behaviour data. The aim is to explore the potential relationships between variables in learning behaviour data and to investigate how learning behaviour characteristics influence learning outcomes. The application of big data analysis results can provide important evidence for postgraduate education quality management and enhance the effectiveness of the postgraduate education quality monitoring information platform.

Funding
This research was supported by the Guangdong Province Education Science "13th-Year Plan" 2020 Annual Research Project: "Special Research on Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era" (202GXJK017).

About the Author
Luo Man (1985.06-), female, Han, Author's Affiliation: South China Normal University, School of Educational Science Doctoral student in Education, Major in Educational Leadership and Management, Research Direction: Educational Leadership and Management.

References

1. Wang, X., Geng, P., Chen, X., Cai, W., & An, H. (2024). A study on the academic innovation ability and influencing factors of public health graduate students based on nomograms: a cross-sectional survey from Shandong, China. Frontiers in Public Health, 12, 1429939.
2. Gates, I. D., Wang, J., Kannaiyan, R., & Su, Y. (2021). Instilling innovation and entrepreneurship in engineering graduate students: Observations at the University of Calgary. The Canadian Journal of Chemical Engineering, 99(10), 2195-2204.
3. Fleith, D. D. S. (2019). The role of creativity in graduate education according to students and professors. Estudos de Psicologia (Campinas), 36, e180045.
4. Schultz, D. M. (2019). Constraints on innovative teaching in British universities: An American perspective. InSight: A Journal of Scholarly Teaching, 14(1), 88-98.
5. Walder, A. M. (2017). Pedagogical Innovation in Canadian higher education: Professors' perspectives on its effects on teaching and learning. Studies in educational evaluation, 54, 71-82.
6. Cai, J. (2018, January). Innovation and Reference of American Applied Talents Training Mode under the Background of" Internet Plus". In 2017 7th International Conference on Education and Management (ICEM 2017) (pp. 865-869). Atlantis Press.
7. Selznick, B. S., Zhang, L., Mayhew, M. J., Bock, C., & Dilmetz, D. (2019). Developing Students' Innovation Capacities: A Comparison between US and Germany. The Three Cs of Higher Education: Competition, Collaboration and Complementarity, 233.
8. Nerad, M. (2020). Governmental innovation policies, globalisation, and change in doctoral education worldwide: Are doctoral programmes converging? Trends and tensions. In Structural and institutional transformations in doctoral education: Social, political and student expectations (pp. 43-84). Cham: Springer International Publishing.
9. Xie, K., Wei, J., Li, Z., Luo, F., Zhou, H., Luo, C., ... & Lu, D. (2022). Reform Thinking and Practice of Innovative Ability Training of Clinical Medicine Professional Master under the Background of Substantial Expansion of Postgraduate Enrollment. Creative Education, 13(10), 3144-3152.
10. Xiang, Y., Ma, Y., Ji, M., & Su, Y. (2024). Interconnected knowledge: Examining the evolution of graduate student innovation ecosystems. Journal of the Knowledge Economy, 15(3), 14036-14075.
11. Yang, B., Bao, S., & Xu, J. (2022). Supervisory styles and graduate student innovation performance: The mediating role of psychological capital and the moderating role of harmonious academic passion. Frontiers in Psychology, 13, 1034216.
12. Horne, L., Soucy, A., DiMatteo-LePape, A., Briones, V., & Wolf-Gonzalez, G. (2024). Reflections of a graduate student team on developing and implementing a transdisciplinary research project: Challenges, recommendations, and lessons learned. Climatic Change, 177(4), 64.
13. Fu, M. (2018, May). Research on the training of graduate students' practice and innovation ability in management science and engineering. In 2018 8th International Conference on Social science and Education Research (SSER 2018) (pp. 344-348). Atlantis Press.
14. Meniailo, V. (2018). Analysis of the current state in innovative research training of PhD students in Ukraine. Advanced education, 101-106.
15. de Souza Fleith, D., & Gomes, C. M. A. (2019). Students' assessment of teaching practices for creativity in graduate programs. Avaliaçao Psicologica: Interamerican Journal of Psychological Assessment, 18(3), 306-315.
16. Pengjun, L. (2024, January). Research on Assessment of Graduate Students' Research and Innovation Ability Based on Distributed Identity Management by Computer Simulation. In 2024 IEEE 7th Eurasian Conference on Educational Innovation (ECEI) (pp. 317-320). IEEE.
17. Logachev, M. S., Orekhovskaya, N. A., Seregina, T. N., Shishov, S., & Volvak, S. F. (2021). Information system for monitoring and managing the quality of educational programs. Journal of Open Innovation: Technology, Market, and Complexity, 7(1), 93.

18. Oseredchuk, O., Drachuk, I., Teslenko, V., Ushnevych, S., Dushechkina, N., Kubitskyi, S., & Chychuk, A. (2022). New approaches to quality monitoring of higher education in the process of distance learning. International Journal of Computer Science & Network Security, 22(7), 35-42.
19. Li, Y., & Zhang, H. (2024). Big data technology for teaching quality monitoring and improvement in higher education-joint K-means clustering algorithm and Apriori algorithm. Systems and Soft Computing, 6, 200125.
20. Kazakhbaeva, M. (2025). MONITORING IN EDUCATION QUALITY CONTROL. Eurasian Journal of Social Sciences, Philosophy and Culture, 5(3), 100-102.
21. Keinänen, M., Ursin, J., & Nissinen, K. (2018). How to measure students' innovation competences in higher education: Evaluation of an assessment tool in authentic learning environments. Studies in Educational Evaluation, 58, 30-36.
22. Si, Y., & Wu, B. (2022). Construction and management method of university information platform based on big data technology. Mobile Information Systems, 2022(1), 7674573.
23. Wang, H. (2023). Teaching quality monitoring and evaluation using 6G internet of things communication and data mining. International Journal of System Assurance Engineering and Management, 14(1), 120-127.