ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Innovative Models of Higher Education Management and Student Training Mechanisms under Big Data Technology

Ruidan Zhang[1,*]

[1] Scientific Research Office, Wuxi Vocational Institute of Commerce, Wuxi, Jiangsu, 214153, China

SUMMARY: *By using a refined K-means algorithm that includes optimized initial centroid selection and lessened distance computations, this research clusters students based on their campus behavior patterns utilizing a sample of 324 students who are enrolled at the XX Vocational College. Associations between these behavior patterns and academic results are then computed through an improved version of the Apriori algorithm. To optimize SVM parameters, a fruit fly optimization algorithm (FOA) is presented to allow early detection of students at academic risk. Main observations indicate that most students spend between 600 and 900 yuan per month, with the average being 789.37 yuan. Internet fees on the campus are mostly 37.64 yuan per month (43.52 percent), but it has been seen that the cost ranges between 9 and 48 yuan. Frequency of bathing is 9-17 per month in 48.77 percent of the sample and the lowest book borrowing group is 73.15 percent of the students who borrow an average of only 6.67 books each. Daily living habits and academic engagement were found as the main determinants of academic performance among the behavioral dimensions evaluated, with spending patterns having relatively low predictive power. It is worth noting that irregular routines seem to result in increased expenditure, implying that lifestyle discipline affects financial behavior too. The suggested model shows high fitting precision and low prediction error, providing a consistent model to be used by vocational college administrators to establish a constant loop of monitoring, early warning, specific intervention, and systematic improvement when it comes to student development.*

KEYWORDS: *Improved K-means algorithm; Improved Apriori algorithm; Association rules; Academic early warning; Higher education management*

1 Introduction

The concept of big data is not merely a technological change, but a sign of the overall change in organization and implementation of knowledge, which adds new impetus to the economic development. The pace of its penetration into literally all spheres of the current society has been fast and extensive and higher education is one of the areas that were impacted the most. Big data platforms have provided higher education institutions with significant capacities of institutional innovation, but this very infrastructure imposes multifaceted and complex pressures on how students learn, relate to others, and identity-forming. Information about an individual who previously could be regarded as personal has become vulnerable to exposure in digitally connected campus settings. The open data networks facilitate the spread of competing value systems, which makes increasing psychological and ideological demands on students,

*zhangruidan@wxic.edu.cn
https://doi.org/10.65102/is2026470
Published: 30 April, 2026

and silently modifies their behavior and attitude. The authority of faculty, which had historically been a pillar of the conventional academic culture, is also being challenged as students can independently access large information resources [1-4]. These converging trends pose real threats to governance of education and student development in modern higher education institutions. Adopting big data is no longer the choice of institutions; it has become quite a necessity. Both educators and administrators need to reconsider and redesign their management strategies, creating structures that respond to the current realities of a data-driven educational environment.

There are numerous perspectives on the use of big data in educational management that scholars have examined. The combination of decision trees and association rules algorithms is used to show in study [5] that the process of mining assessment data becomes significantly more effective, and thus, adds more scientific value to the institution governance and contributes to developing digital construction. To address another group of issues, study [6] aims at the long-standing vulnerabilities of graduate education, i.e., poor cohort identity, segmented information flow, and unsychronized administrative supervision, and offers a management framework based on big data, which can integrate online resources, enhance administrative responsiveness, and better meet the needs of students. Study [7] has a more holistic theoretical perspective that claims that profound integration between big data and managerial practice may enhance foresight, customize student support, and enhance more interactive institution relationships, eventually promoting the modernization of higher education governance. Technically, study [8] gathers the data on the campus with the help of IoT devices and analyzes it with the help of K-means clustering, principle component analysis algorithms and Apriori algorithms. The aim is to unearth actionable insights in heavy data streams without compromising privacy, which will lead to the digitization of teaching administration. Decision tree modeling is integrated with distributed computing platforms like Hadoop and Spark to reinforce the effectiveness of a dedicated student management system, as per study [9] . The enhanced decision tree model achieves the level of 92% in terms of behavioral prediction accuracy and the associated smart warning and personalized recommendation systems achieve tangible improvements in performance efficiency with the ability to accurately intervene and provide individualized learning support.

In student training, big data technology is optimized and innovated in multiple dimensions, such as training objectives, teaching mode, teaching evaluation, etc., to promote the innovation of student training mechanism. Literature [10] uses big data technology, commercialization and innovation strategies to improve the management of higher education, and finds that commercialization means such as technology transfer and industrial alliances not only create economic benefits, but also promote the flow of knowledge and enhance the social impact of the institution. Literature [11] proposed a framework centered on "interdisciplinary knowledge network - problem-solving ability - human-computer collaborative practice". By leveraging big data and artificial intelligence technologies, it aims to shift higher education from a "subject-oriented" model to a "competence-oriented" one, cultivating comprehensive talents with critical thinking, interdisciplinary skills, and human-computer collaboration capabilities. Literature [12] constructs an evaluation system through hierarchical analysis to diagnose the deficiencies of the current model in terms of professional characteristics, student demand satisfaction, etc., and introduces a big data system to optimize the cultivation path to provide data-driven decision-making support for the cultivation of high-level talents with international vision and competitiveness. The investigation of big data-based methods of talent development and teaching quality has produced various educational outcomes. By restructuring the curriculum, enhancing the industry-academic partnership, and providing personalized training in the form of big data tools, study [13] develops an intelligent teaching optimization strategy that explicitly

targets the creation of digitized graduates that can meet the current needs of the industry. Study [14] also follows the same student-focused objective by creating advanced models that will inform the allocation of resources and develop customized learning techniques. The adaptive environments that are created are said to enhance the level of engagement of students as well as the overall effectiveness of learning and to strengthen student management practices. The research of [15] evaluates the issue by structuring it with three dimensions of analysis that include teaching effectiveness, behavioral patterns, and competence development. Based on the closed-loop logic of planning, execution, assessment, and feedback, the research creates a comprehensive knowledge evaluation model that has several assessment modes, three interrelated dimensions, and four consecutive procedural steps. In a data reduction strategy, study [16] uses random forest algorithms to monitor graduate quality, reducing unnecessary variables based on feature ranking and training classifier to measure individual capability. It provides higher education institutions with a technical path for assessing talent cultivation results more accurately. The research of [17] expands this objective to outcome-based education, suggesting a model of assessment that is based on both big data and artificial intelligence. The framework covers the instructional process and employment outcomes and increases evaluation objectivity by 38 percent and enhances the mean alignment between graduate competencies and job requirements by 21 percent.

The research goes on in three interrelated steps. Firstly, student educational data are collected and pre-processed, and afterwards, the traditional K-means algorithm is optimized to perform better in two specific ways: choosing initial centroids in a more principled manner, and decreasing the overall number of distance calculations needed. This improved algorithm is subsequently used to group students in terms of behavioral patterns related to their consumption behavior, daily routines and academic participation. The next stage is the focus on how these behavioral profiles relate to academic performance. To achieve this goal, association rule mining is defined, and the Apriori algorithm is reinforced using a pruning process that reduces unnecessary candidate itemsets to provide more clear and credible associations between student behavior and scholastic achievement. Stage three focuses on early warning. An SVM parameter optimization plan is created based on an enhanced FOA with an adaptive step size mechanism that enables the model to optimize its parameters more sensitively. This setup allows identifying students at academic risk on time and re-positioning the institutional management from a passive and reactive stance to one where anticipation and proactive intervention take the lead.

2 Data mining and processing for student education management

2.1 Pre-processing Methods for Student Behavior Data

2.1.1 Data sources

This work is based on the empiricism behind this research, which is the behavioural data produced by a student population of XX Vocational College where the main source of data will be the campus card purchase records. Raw data sets were three, viz 31,024 library borrowing records, over 23 million card-based consumption transactions and about 3.85 million academic grade entries covering the full student population. The borrowing data set contains both information on student details and bibliographic data, including the name of the borrower, book author, publisher and book identifier as well as timestamps of loan and return events and a borrowing identification number. There are thirty five attributes in total but the analysis uses just six of them: student card number, name, outstanding balance, borrowing date, return date

and book title. Consumption records by campus cards record a larger transactional image, containing student identifiers, dates and values of every transaction, recharge events, and point-of-sale locations in 13 attributes. Five of them are used analytically, i.e., student card number, name, transaction time, transaction amount, and remaining balance. The grade records are the most attribute-rich dataset, initially containing 50 fields covering subject names, credit values, the designation and level of remedial courses, course categories, student classification, and instructors. Since the analysis is one of several in this paper, only six attributes are kept: student card number, name, grade, credits, and academic year, and semester.

2.1.2 Data processing

(1) Data Mining

Data mining is the use of some technology to find the laws and characteristics of the data, using data features to help people make some scientific decisions and judgments. The features obtained from data mining have an important impact on people's lifestyle and level, and managers can make scientific decisions and predictions based on the data features to improve students' course-passing rates in vocational colleges.

(2) Data Cleaning

The process of data cleaning includes processing missing values, eliminating noise data, and eliminating inconsistencies between data. According to whether the variables are missing or not the data are divided into complete variables and non-complete variables, in which complete variables are those that do not contain missing values and non-complete variables are those that contain missing values.

(3) Data integration

The process of data integration is bringing together records that have been pulled out of various heterogeneous sources into a single, consistent framework. During this stage, it is important to pay close attention to the naming conventions as well as formatting requirements of every contributing source. The conflicting values are discovered and dealt with systematically and on the case-by-case basis whereas the redundant entries are found and removed so as to not compromise the integrity and analytical usefulness of the combined dataset.

The relevance measure to determine whether an attribute is redundant is:

$$\gamma_{A,B} = \frac{\sum (A - \bar{B})(B - \bar{B})}{(n-1)\sigma_A \sigma_B} \tag{1}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the mean values of A, B respectively, and σ_A and σ_B are the standard deviations of the attributes A and B respectively:

$$\sigma_A = \sqrt{\frac{\sum (A - \bar{A})^2}{n-1}} \tag{2}$$

$$\sigma_B = \sqrt{\frac{\sum (B - \bar{B})^2}{n-1}} \tag{3}$$

With $\gamma_{A,B} > 0$, there is a positive correlation between the A, B properties. In such a situation, the A value increases with the B, and higher values are associated with a greater likelihood of one property implying the presence of another. On reaching an adequately high level of this value, either of these two attributes can be considered redundant and eliminated

without losing any meaningful information.

Under the $\gamma_{A,B} < 0$ condition, the interaction between the A,B properties is negative. In this case, the A value changes in the opposite direction to B, indicating a suppressive trend whereby the existence of one property decreases the probability of the other being present.

When the $\gamma_{A,B} = 0$ is true, the A,B characteristics have no systematic correlation with each other and are viewed as mutually independent.

(4) Data transformation

In the context of data transformation, normalization is the most popular technique used to convert the raw data into one that can be mined. It works by rescaling the attribute values on a particular range so that they fit into a given numerical interval.

1) Minimum-maximum normalization

For a given numerical attribute A, $\left[\min_A, \max_A\right]$ is the interval of values before specification $\left[new\_\min_A, new\_\max_A\right]$ is the interval of values after specification, and minimum -max specification specifies the value v of A as v' according to the following equation:

$$v' = \frac{v - \min_A}{\max_A - \min_A}\left(new\_\max_A - new\_\min_A\right) + new\_\min_A \tag{4}$$

2) Zero-mean specification

With respect to any numerical attribute A, the values of $\bar{A}$ and σ_A represent the average and standard deviation of A respectively. In zero-mean normalization, every original value v of A is normalized and transformed into a transformed value v' via this equation:

$$v' = \frac{v - \bar{A}}{\sigma_A} \tag{5}$$

3) Fractional calibration specification

Given a numerical attribute A, $\max|A|$ is the maximum absolute value of A, and j is the smallest integer that satisfies $\frac{\max_{|A|}}{10^j} < 1$. Decimal scaling normalization transforms each original value v of A into a rescaled value v' through the following equation:

$$v' = \frac{v}{10^j} \tag{6}$$

(5) Data Statute

The data statute can also be called data reduction or data simplification. For large data sets, the data statute can be expressed through data generalization, the representation is much smaller than the original data, but still close to maintain data integrity, so that after the statute of the data set on the analysis of more efficient, and produce the same analysis results. The data statute has two main categories: attribute statute and record statute.

(6) Data Analysis

Raw data for data preprocessing after the analysis of data, data analysis in a variety of ways are mainly divided into four major modules, namely, frequent pattern mining, cluster analysis, classification, machine learning and so on.

(7) Data preprocessing

Kettle is written in Java based on open source, provides a graphical interface, can be connected to any database and other features, the tool is able to screen the data for attributes, data diversion operations. Data extraction process shown in Figure 1. The use of the software can be extracted from the database or storage files to the required data, simplify the data extraction steps. Kettle has two script files are transformation and job.

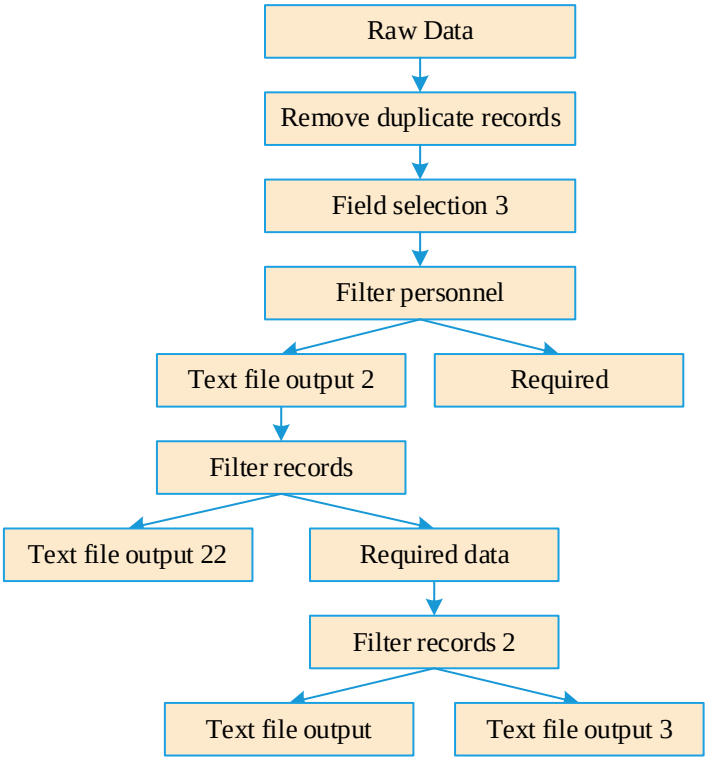

Figure 1: Data extraction process

2.2 Cluster analysis method for students' daily behavior

2.2.1 K-means algorithm

The K-means algorithm has traditionally used the sum of squared errors (SSE) as the main measure of clustering quality. In this context, D_i is the ith cluster and x is a sample point of D_i, d is the distance between two objects, and k is the total number of clusters available.

$$SSE = \sum_{i=1}^{k} \sum_{x \in D_i} d(D_i, x)^2 \tag{7}$$

With a given value of k, a lower SSE is indicative of a better result of clustering. When k grows, samples are divided into finer pieces, intra-cluster cohesion goes up, and the sum of squared errors goes down accordingly.

2.2.2 Improved K-means algorithm

The approach to choosing initial centroids used in this study is based on the density. Consider a given data $X=\{x_1,x_2,\cdots,x_n\}$ with x_i being a single sample point, k representing the number of clusters, n the total number of samples and N a relative count parameter. Firstly, the density $D(x_i)$ of the sample points in the dataset and the density threshold of the dataset Density threshold are obtained, and the two are compared, and then the sample points are assigned to the high-density sample set H or the low-density sample set L, and the samples with a density greater than the threshold are classified into the H sample set, and the samples with a density samples smaller than the threshold are judged as outliers and are categorized into the L sample set. To make sure that two or more initial centroids are not selected in the same cluster region, spatially scattered selections are preferred. Both representativeness and the designation of the sample point with the highest density value among set H as the first initial centroid and the removal of set μ_1 are guaranteed simultaneously. , and then removes it from the sample set H; The second initial center of mass μ_2 is selected such that it satisfies that the distance $d(\mu_1,\mu_2)$ between μ_2 and μ_1 is greater than a threshold value r and μ_2 should have the maximum density at this point, and so on until k initial centers of mass are selected.

Any two arbitrary samples x_i and x_j are measured as a distance based on the Euclidean distance formula:

$$d(x_i,x_j)=\sqrt{(x_i-x_j)^2},\quad i=1,2,\cdots,n,\ j=1,2,\cdots,n \tag{8}$$

The density of any sample point x_i in the dataset X is denoted as:

$$D(x_i)=N\left\{x\middle|d(x,x_i)<r,\ x\in X\right\} \tag{9}$$

Calculating the average density of the dataset yields:

$$avgD(x_i)=\frac{\sum_{i=1}^{n}D(x_i)}{n} \tag{10}$$

Calculating the dataset density threshold, where c is a constant, yields:

$$Density\ threshold=c\times avgD(x_i)=c\times\frac{\sum_{i=1}^{n}D(x_i)}{n} \tag{11}$$

For the problem of distance calculation, the traditional K-means algorithm is cumbersome, this paper optimizes the algorithm by reducing the number of distance calculations. Firstly, the Euclidean distance $d(x_i,\mu_1)$ from all sample points to the first initial center of mass μ_1 selected after optimization, and the distance $d(\mu_1,\mu_2)$ between μ_1 and μ_2 are calculated, according to the nature of space cubic geometry, if $2d(x_i,\mu_1)\le d(\mu_1,\mu_2)$, then $d(x_i,\mu_1)\le d(x_i,\mu_2)$, i.e., the sample point x_i is closer to the center of mass μ_1, then there is no need to compute $d(x_i,\mu_2)$, the sample point does not belong to the cluster set where μ_2

is located; and so on, to select which center of mass point x_i is closer to, and will be farther away from the sample point The initial center of mass that is farther away from the sample point is screened out, and only the initial center of mass that is closer to the sample point is considered until the k th initial center of mass. In short, the method only requires the calculation of the distances between the k initial centers of mass, the distances from all sample points to the first initial center of mass μ_1, and the distances between the sample points and the initial centers of mass that are closer to them. According to this method, the samples can be divided into clusters that are closest to them, which greatly reduces the number of calculations of the distances between the samples and the initial centers of mass and improves the efficiency of the algorithm.

The specific flow of the improved K-means algorithm is shown below:

Input: dataset $X = \{x_1, x_2, \cdots, x_n\}$, the number of clusters k, and the density threshold Density threshold.

Output: Cluster division and initial center of mass.

(1) Calculate the density $D(x_i) = N\{x | d(x, x_i) < r, x \in X\}$ of the sample points x_i in the dataset X, as well as the density threshold of the samples $\textit{Density threshold} = c \times \frac{\sum_{i=1}^{n} D(x_i)}{n}$;

(2) Compare the sample density $D(x_i)$ with the density threshold Density threshold, if $D(x_i)$ is less than the Density threshold, the sample is determined as an outlier and classified into the low-density sample set L, and vice versa, the sample is classified into the high-density sample set H;

(3) Select the sample point with the highest density in the sample set H as the first initial center of mass μ_1 and move it out of the sample set H;

(4) Select the second initial center of mass μ_2, where $d(\mu_1, \mu_2) \geq r$ and $D(\mu_2) = \max\{D(x) | x \in X - \{\mu_1\}\}$;

(5) Repeat step (4) until k initial centers of mass are selected;

(6) Calculate the distances between the k initial center of mass points and the distances from all sample points to μ_1;

(7) Compare the size of $2d(x_i, \mu_1)$ with $d(\mu_1, \mu_2)$, and if $2d(x_i, \mu_1) \leq d(\mu_1, \mu_2)$, $d(x_i, \mu_1) \leq d(x_i, \mu_2)$, and x_i is nearer to μ_1, then x_i does not belong to the clustering category where μ_2 is located;

(8) and then calculate the distance between the sample and the next initial center of gravity has not been compared to select the sample point to which center of gravity point is closer to the comparison of the initial center of gravity farther away from the sample point will be excluded;

(9) Repeat step (8) up to the k th initial center of mass, and so on until all samples are classified into the closest clusters.

2.3 Methods for analyzing the correlation between student behavior and achievement

2.3.1 Introduction to association rules

The related concepts of association rules are as follows:

(1) The term set is represented by the formula:

$$I = \{i_1, i_2, i_3, \cdots, i_n\} \tag{12}$$

where I is then denoted as the item set, which contains all the items i_j, and i_j denotes a specific item.

(2) The database is shown in the following equation:

$$D = \{t_1, t_2, t_3, \cdots, t_n\} \tag{13}$$

where t_j is a purchase record that consists of several items in I, i.e., $t_j \subseteq I$ is called a transaction, and all transactions make up the database D.

(3) Rules

In the association rule, if $X \subseteq I, Y \subseteq I$, can be introduced by $X \Rightarrow Y$ and $X \neq \varnothing, Y \neq \varnothing, X \cap Y = \varnothing$, such as this kind of association rule, X, Y are a subset of I, and there is no intersection between them, when can be introduced by X, Y, X and Y will be established between the association.

(4) Support: the ratio of all transactions in a dataset in which a particular combination of products occurs simultaneously compared to the overall number of transactions. The support formula used in this research is given as:

$$Support(X, Y) = P(X, Y) = \frac{number(XY)}{number(AllSamples)} \tag{14}$$

Prior to analyzing the relationship between X and Y, it is important to ensure that the rate of their co-occurrence in the database is at least some minimal level because this condition forms the basis of statistical reliability of any rule that can be formulated according to the data.

(5) Confidence is a measure of the conditional probability that an item can be found based on the fact that another has been observed previously and it is indicative of the strength of direction of an association between two pieces of data.

After the support of X and Y reaches the minimum support threshold, the existence of a true associative relationship between them becomes a matter of consideration that should be addressed with confidence. The appropriate formula is presented below:

$$Confidence(X \Rightarrow Y) = P\left(\frac{X}{Y}\right) = \frac{P(XY)}{P(Y)} \tag{15}$$

To sum up, there are two consecutive steps in association rule mining. In the first step, it is necessary to identify all high frequency item sets that satisfy the minimum support condition, and in the second one, it is required to derive meaningful association rules out of those high frequency item sets.

2.3.2 Improved Apriori algorithm

(1) Introduction of Apriori algorithm

The core idea of Apriori algorithm is to derive the combination of K items from the combination of $K+1$ items, which relies on the theory that if the combination of $K+1$ items is a high-frequency combination, the subset of K items must also be a high-frequency

subclass. The group in the K term is taken as a subclass of the $K+1$ term group, and according to its cross-combination, all the candidate term groups that may be high-frequency term groups are listed, and then the subset of K terms of the $K+1$ term combination that is not included in the K term combination is deleted, and finally their support is counted from the original dataset individually, and those candidates that do not meet the standard are removed that is, we get our $K+1$ term high-frequency combinations.] term high-frequency combination.

The core process of Apriori algorithm:

1) Connection step: the Apriori algorithm is based on the a priori nature of frequent itemsets as a theoretical basis, so we can obtain C_k by the cross connection of L_{k-1}. Let the items in the two crossed L_{k-1} be I_j, I_k respectively, and both of them are arranged in dictionary order, when the items in L_{k-1} are cross-connected, it is required that the first $k-2$ items of I_j, I_k are the same $\left(I_j[1]=I_k[1] \wedge I_j[2]=I_k[2] \wedge \cdots \wedge I_j[k-2]=I_k[k-2]\right)$, then I_j, I_k can be concatenated, in order to make sure that the concatenation result is not repeated $I_j[k-1]<I_k[k-1]$.

2) Pruning step: C_k generated by the connection step, we need to prune C_k to remove the infrequent terms to reduce the amount of computation.

(2) Improvement of Apriori Algorithm

The original Apriori algorithm, after cross-generating C_k from L_{k-1}, needs to compare whether the subset of $k-1$ in C_k exists in L_{k-1} for cropping, and then it needs to check the subset of $k-1$ of all the items in C_k with the subset of L_{k-1}, which generates a large amount of computation, and the improvement of Apriori algorithm in this paper is the optimization in this aspect. The improvement of Apriori algorithm in this paper is the optimization in this aspect.

In order to reduce the amount of computation in the process of checking the $k-1$ subsets of all the terms in C_k against the subsets in L_{k-1}, we change the way of comparison from the original way of selecting a term in C_k and decomposing its subset to check it against the subset in L_{k-1} to the way of checking the subset of L_{k-1} against the subset of all the terms in C_k, which will generate a large amount of computation, and this improvement of the Apriori algorithm in this paper is the optimization of this link. The subset kerneling of all terms is shown below in its formulaic presentation:

Variable description: $|C_k|$ denotes the number of subsets of C_k, $|L_{k-1}|$ denotes the number of subsets of L_{k-1}, $(C_k)_i$ denotes the subset of the i th term in C_k, and $(L_{k-1})_j$ denotes the subset of the j th term in L_{k-1} and it can be known that the number of subsets produced by each term in C_k is C_k^1. So the computational amount of the pruning step in the original Apriori algorithm can be expressed as:

$$\sum_{(C_k)_1}^{(C_k)_{|C_k|}} C_k^1 \left|L_{k-1}\right| \tag{16}$$

The improved calculation is:

$$\sum_{(L_{k-1})_1}^{(L_{k-1})_{|L_{k-1}|}} \left|C_k\right| C_k^1 \tag{17}$$

The whole pruning process is: check the subset in L_{k-1} against the subset of all the terms in C_k respectively, and assign a value to that C_k term according to the match, and so it is divided into two cases to be discussed as follows:

a. $|L_{k-1}|-C_k^1+1>C_k^1$

The above equation is expressed as when the matching process between the items in L_{k-1} and C_k, if the number of non-compliant items exceeds the number of each subset of items in C_k^1, we use the number of matches judged to be compliant to eliminate the non-compliant items in C_k in the pruning process.

b. When $|L_{k-1}|-C_k^1+1<C_k^1$, then we use the number of judgmentally non-conforming matches to eliminate the non-conforming items in C_k during the cropping process.

3 Analysis of data on student education management

3.1 Analysis of student behavior

3.1.1 Behavioral analysis of student consumption levels

The expenditure of students on campus can be divided into two large groups. The first one includes card-based purchases in dining halls, supermarkets, and other campus locations with card-payment capabilities. The second one involves both online orders and cash payments in off-campus locations without passing through the campus card system. Every instance when a card payment is made, an associated record is automatically added to the campus card database. Since the amounts of individual transactions differ significantly, it is necessary to aggregate these records by student numbers and add up the registered amounts to obtain a representative image of the total amount spent by each student in a certain timeframe. These grouped data were averaged monthly and plotted as a distribution graph as shown in Figure 2. The plot shows that most students spend between 600 and 900 yuan on average monthly.

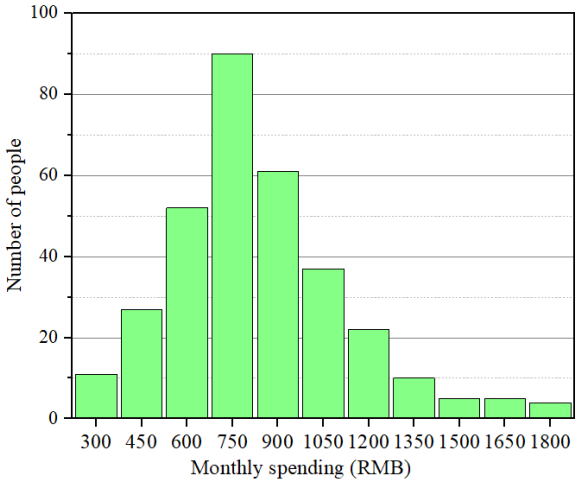

Figure 2: Distribution of students' average monthly spending

As per the analytical goals of the given research, the patterns of student consumption are divided into three levels: high, medium, and low. The value of the number of cluster centers k is then set to 3. Experimental trials show that the lowest sum of squared errors is achieved when

the random seed is set to 5 indicating the most stable and reliable clustering result. The obtained average monthly consumption distribution is shown in Figure 3. Cluster analysis will sort all sampled students into one of three groups, which are 11.73, 62.65 and 25.62 percent in clusters 1 through 3 respectively. These groups represent average monthly expense amounts of 409.45 yuan, 789.37 yuan, and 1,271.04 yuan on average during the timeframe between 2023.6 and 2025.6, which can be directly mapped to low, medium and high consumption categories respectively. Most of the students therefore lie in the medium consumption category with an average of less than 800 yuan per month.

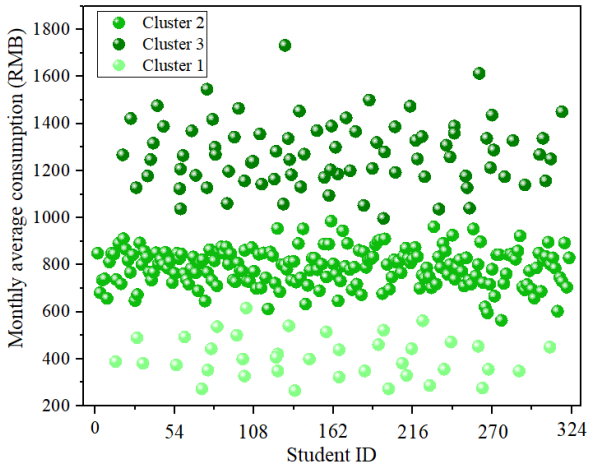

Figure 3: Student's average monthly spending cluster results

The results of clustering are converted into discrete consumption tiers based on the student and one of these classifications is shown in Table 1. The statistics show that 16.67 percent of the student population belongs to the low consumption category. This result has practical significance to the welfare management of students. These findings could be used by faculty advisors and student affairs personnel to inform them to take proactive steps towards contacting the students in this tier so that they can have a better picture of their unique situations. These individuals whose financial suffering is supported by the information might now be given priority consideration when awarding scholarships or assigning work-study opportunities. The process of making such choices based on empirical evidence instead of subjective opinion enables institutions to make their decisions faster and more accurately so that the support tools can get to the students that actually need help.

Table 1: The consumption level of some students

Student ID	level of consumption
2024012278	High level
2024012813	High level
2024013196	Intermediate level
2024012636	Intermediate level
2024014029	Intermediate level
2024012126	High level
2024012601	Intermediate level
2024012774	Intermediate level
2024013340	Lower level
2024013085	Intermediate level
2024012138	Lower level
2024012860	High level

3.1.2 Analysis of students' access to the Internet

The college network center monitors the day-to-day internet activity of students and through analysis of their online histories, it can be determined how they use the internet. It is usually believed that there is a positive correlation between the monthly payments made by students to access campus internet services and the time they spend online. In order to investigate this connection, information related to the consumption of one-card on campus by students is obtained and studied. This data shows what distribution of average monthly internet spending by students at campus looks like and is shown in Figure 4. These observations reveal that most of the students have their monthly internet charges not exceeding the interval [9, 48].

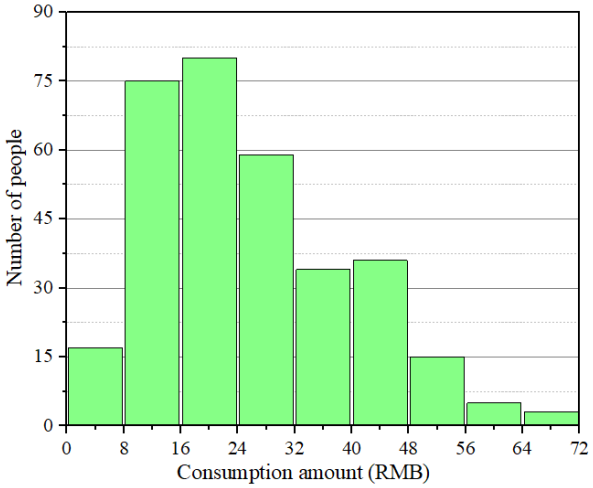

Figure 4: Distribution of Monthly Average Cost of Students' Campus Network

Through the process of K-means clustering with respect to the average monthly campus internet usage, it is possible to classify students according to their spending behaviors as indicated in Figure 5. The three different categories of students have been created depending on their average monthly spending on the internet. Category 1 is made up of those who have low internet consumption with an average monthly expenditure of 13.30 yuan per student which amounts to 31.48% of the total number of students. Category 2 is comprised of individuals who have moderate consumption with a mean monthly expenditure of 37.64 yuan, which is 43.52% of the student body. Lastly, Category 3 contains the students who have high internet consumption, which averages to 70.28 yuan per month, and this category comprises 25 percent of all the students. College administrators need to pay closer attention about monitoring the daily lives of people belonging to the higher consumption category and give advice and assistance to students experiencing difficulties so that they can be healthy and develop in a balanced way.

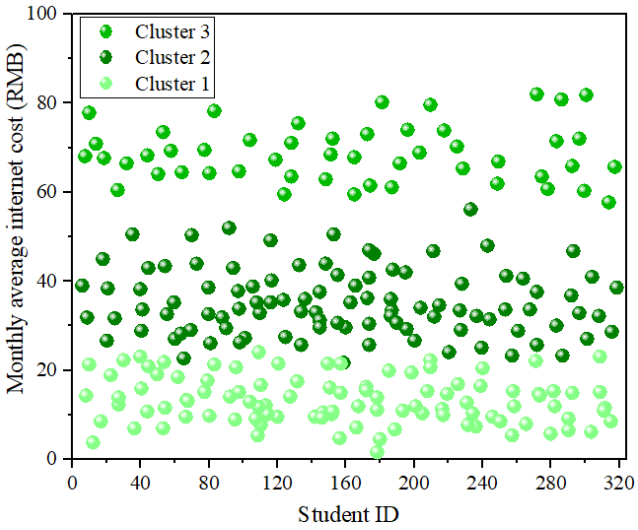

Figure 5: Average monthly consumption clustering results of campus network

3.1.3 Consumer behavior analysis of student bathhouses

The information of student bathing consumption is based on filtering and processing the campus card swipe data whereby the quantity of bathing time per student is determined using data transformation. K-means clustering was used to classify the students depending on the frequency of bathing as illustrated in Figure 6. Three groups of students were created with each group reflecting a distinct level of bathing frequency, low, medium, or high. The first group, category 3, is the group of students with the lowest level of bathing frequency, an average of 47-7 baths per month, which amounts to 38.27 percent of the student population. Category 2 refers to those with a moderate frequency, 9-17 baths per month, which is 48.77 percent of the overall students. Lastly, the category 1 contains those students who bathe most frequently with an average of 18-22 baths per month and that group constitutes 12.96 percent of the entire student population.

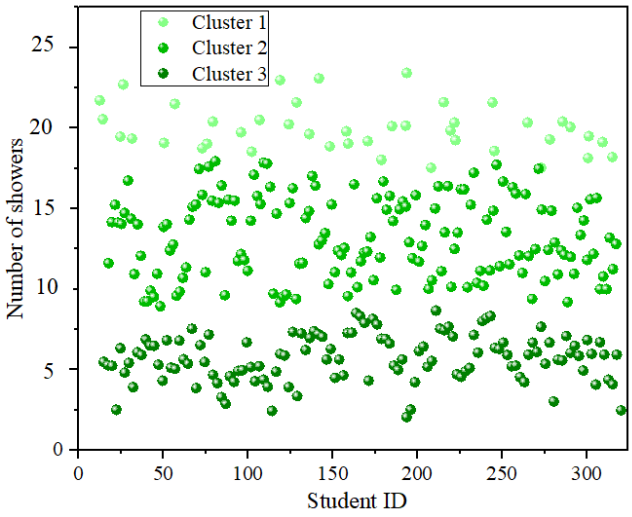

Figure 6: Student shower frequency clustering results

3.1.4 Analysis of Students' Library Borrowing Behavior

College libraries provide students with access to vast collections of books covering general

interests as well as those related to specific disciplines, and borrowing behavior will naturally be a mix of personal interest and academic need. Three-cluster classification of the cluster analysis performed on the library borrowing data of 324 sophomore students results in three distinct groups whose distribution is shown in Figure 7. The first and biggest group consists of students with the least borrowing actions, with an average of 6.67 volumes per person, and they represent 73.15 percent of the total population. The second group is in the middle with the members borrowing an average of 15.84 books per person and they represent 21.29 percent of the total. The third group, which is only 5.56 percent of students, exhibits significantly greater involvement with library materials with an average of 31.86 books per person. Combined, these statistics indicate a clear-cut image: most students use the library in a low-intensity manner, whereas highly active and regular users are a tiny part of the student population.

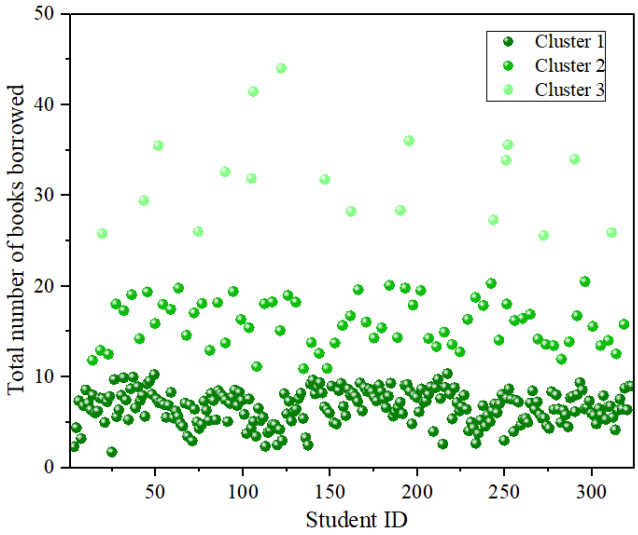

Figure 7: Student book borrowing cluster results

3.2 Student Behavior Habits Association Rule Mining Analysis

3.2.1 Experimental Analysis of Improved Apriori Algorithm

The purpose of the comparative experiment is to compare the computational efficacy of the suggested Apriori variant to assess the performance of students on the basis of their behavioral and academic records at the same threshold settings used among all the other methods tested. Four algorithms are considered simultaneously: the ordinary Apriori, the compressed-Apriori variant, the Apriori-hash variant, and the algorithm introduced in this paper. It is compared to three open-source datasets, namely Groceries, Movies, and Adult, where execution time is the main measure. Comparative findings are summarized in Table 2. Through multi-layered compression of the iterative scanning process, the proposed algorithm makes its efficiency savings by significantly reducing redundant scanning of the transaction set and minimizing the load on the CPU during the process of mining. In all the three benchmark datasets, the method always performs faster than its competitors. Compared to the Apriori-hash algorithm alone, the proposed system takes less time to run (by a factor) of 3.35 - 4.91, which is significantly meaningful in the mining context of associations between the behavior of students and their academic performance.

Table 2: Time-consuming results of four algorithms on three data sets

Algorithm	Groceries	Movies	Adult
Apriori	73.59s	33.82s	17.86s
Apriori-compression	67.35s	22.98s	11.64s
Apriori-hash	15.57s	20.67s	16.89s
Our algorithm	3.28s	4.21s	5.04s

3.2.2 Analysis of the correlation between students' behavior and performance

The cluster analysis results according to the student behavior data are applied to determine different clusters of students and record the distinguishing features of each cluster. The summarized group profiles and related category labels form the main input of the next association rule mining step. Table 3 is a detailed presentation of every behavioral dimension with the respective category labels of each dimension.

Table 3: Summary of Different Types of Labels and Category Labels

Student behavior	Class	Category Tags	Type Label
Consumer Behavior Indicator Label	1	Low living expenses, stable and regular consumption, with virtually no additional expenditures	Hypoglycemic type
	2	Exorbitant expenses, significant fluctuations in single transactions, frequent takeout consumption, and high additional costs	High consumption type
	3	Low living expenses, stable and regular consumption, and minimal additional expenditures	Frugal type
	4	Moderate expenditure, moderate consumption frequency, and regular consumption patterns	Moderate type
	5	Low consumption frequency, low expenditure on three meals, and high additional expenses	Low consumption campus
Regular Lifestyle Behavior Indicator Label	1	Irregular meal patterns, reliance on takeout, minimal morning wake-up, and prolonged internet usage	Very irregular type
	2	Regular meal schedule, occasional takeout ordering, frequent early morning wake-up, and short internet usage time	Regular pattern
	3	irregular meal patterns, infrequent takeout consumption, infrequent early morning wake-up, and moderate internet usage	Regular pattern comparison
	4	Irregular meal patterns, infrequent early morning wake-ups, frequent takeout ordering, and prolonged internet usage	Irregular type
Learning Behavior Indicator Label	1	Average book borrowing volume, frequent library visits, and prolonged self-study time	Effortful type
	2	Frequent borrowing of books, moderate library visits, and short self-study periods	Not trying hard enough
	3	Frequent borrowing of books, high library visit frequency, and moderate self-study duration	More diligent type
	4	Less book borrowing, fewer library visits, and reduced self-study time	Not striving type

Academic achievement is the core of the students education experience, where grades are the most straightforward and commonly known indicator of educational achievements. The

mark of one course denotes how much a student has understood the material taught but the total GPA provides a more comprehensive picture of the student's overall academic performance in every course. The combination measure is calculated based on the weighted average of all course grades, which are adjusted based on the course credits and the scale goes between 0 to 5. Marks of 4 to 5 are considered outstanding, marks of 3 to 4 are rated as good, 2 to 3 are acceptable, 1 to 2 are passing and any score under 1 is considered unsuccessful. To produce a joint set of data that will be examined later, both student behavior and academic performance are combined and examined together. Findings of this joint analysis are given in Table 4.

Table 4: Results of Student Behavior Data and Grade Data

Student ID	Consuming behavior	Regular lifestyle behaviors	Learning behavior	GPA level
2024012278	Moderate type	Regular pattern comparison	Not trying hard enough	Good
2024012813	High consumption type	Irregular type	Not trying hard enough	Secondary
2024013196	Moderate type	Regular pattern	Effortful type	Outstanding
2024012636	Moderate type	Regular pattern comparison	Not trying hard enough	Good
2024014029	Frugal type	Regular pattern	More diligent type	Good
2024012126	High consumption type	Irregular type	Not striving type	Secondary
2024012601	Low consumption campus	Irregular type	Not trying hard enough	Secondary
2024012774	Moderate type	Regular pattern	Effortful type	Outstanding
2024013340	High consumption type	Irregular type	Not striving type	Bad
2024013085	Hypoglycemic type	Regular pattern	Not trying hard enough	Good
2024012138	Moderate type	Regular pattern comparison	Not trying hard enough	Good
2024012860	Moderate type	Regular pattern	Effortful type	Outstanding

The analysis to determine the association rules of student behavior and achievement data was performed through the enhanced Apriori algorithm to investigate the connection between students and their habits and their academic performance. The lowest support value was taken as 0.25 and the lowest confidence value as 0.5. In this way, 37 association rules were identified, and the network diagram representing these rules is illustrated by Fig. 8. Association rule mining between students behavior and achievement is presented in Table 5.

Of the found rules, two of them were especially notable as high-performance indicators: Rule #1 and Rule #2. The first rule says: Regular, hard-working behaviors result in high performance, which indicates that behaviors like regular meals, takeout at times, frequent early risings, limited internet use, occasional borrowing of books, and extended library study are linked to good grades. Rule #2: Greater regularity and effort result in better performance, implying that those with less regular eating habits, less takeout, more irregular early mornings, medium internet use, less book borrowing, higher frequency of visiting the library and longer study duration are likely to earn better grades. It can be inferred based on either of these rules that academic performers usually follow a fairly organized lifestyle, get up early, browse the

internet moderately, and take their meals at the cafeteria with a few takeouts, and read their textbooks in the library for an extended period without borrowing books regularly.

Two additional guidelines establish behavioral trends that can be linked to good instead of excellent performance. In this tier, students have relative weaknesses in some aspects of lifestyle compared to the most successful students in their class, but their long-term academic performance is enough to ensure that they remain competitive in terms of grades.

Another five rules, number 5-9, are associated with moderate academic results. In all these rules, lack of involvement in studies is identified as the major cause of poor performance, and inconsistent daily routine is identified as the secondary contributor to this problem. On the other hand, consumption behaviors have very limited meaningful relationship with grade performance in this population. Those students who were in such patterns could gain advantages of developing more disciplined study routines and learning practical lessons out of the habits of more successful classmates.

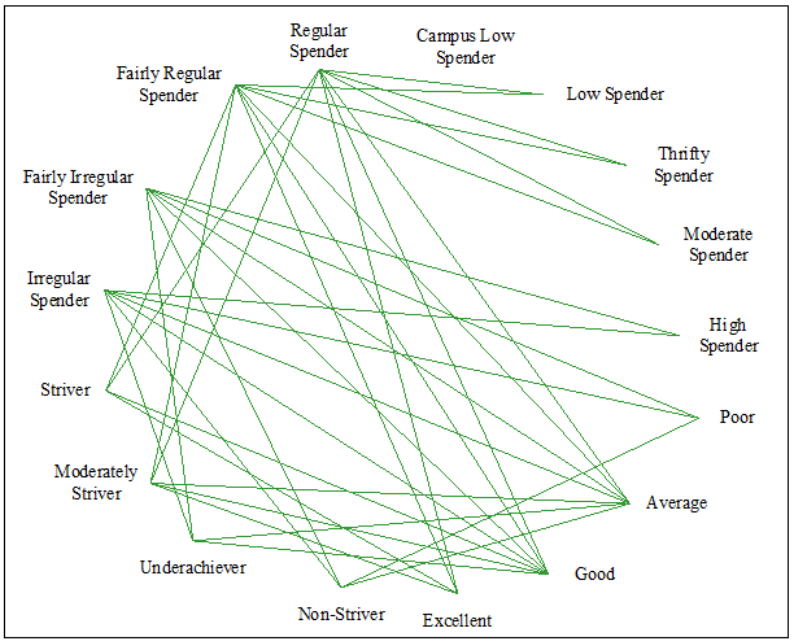

Figure 8: Association rule network diagram

Table 5: Association Rules Mining of Student Behavior and Academic Performance

Number	Rule	Support	Confidence
1	Regular pattern, diligent type => excellent	0.4429	0.8657
2	Regular and diligent => Excellent	0.4262	0.8446
3	Regular pattern, more diligent type => good	0.5164	0.7348
4	More regular, more diligent => Good	0.5158	0.7296
5	Regular pattern, not enough effort => Medium	0.3513	0.7446
6	Less regular, less diligent => Moderate	0.4335	0.9146
7	Moderate, less regular, not enough effort => Moderate	0.2289	0.7839
8	Low-calorie campus life, lack of effort => Moderate	0.2131	0.7723
9	Frugal, regular, not hard enough => Medium	0.4337	0.7795
10	Hard=>Not Effortful Type	0.5442	0.9194
11	regular=>irregular	0.5534	0.8872
12	Meaning=>inadequate effort type, irregular type	0.5322	0.8846

4 Early warning methods for students' academic risks and analysis of early warning effects

4.1 Methodology for early warning of students' academic risk

4.1.1 SVM hyperparameters

The underlying academic risk warning model dataset has combined four sources, namely, student grades record, library admission, book borrowing history, and campus card transaction information. Since this joint dataset is not linearly separable, a non-linear SVM classification system is used and the Gaussian kernel is chosen to perform the mapping operation. There are two parameters that regulate the behavior of this model in very important ways. The first one, the penalty coefficient γ, is used to regulate the severity of the penalty imposed on the misclassified examples when training, which can be seen as a trade-off between margin width and classification error tolerance. The second one, gamma, is an internal parameter of the Gaussian kernel that defines how the initial low-dimensional data is distributed over the high-dimensional feature space after the kernel transformation. Both of these parameters play significant roles in the overall generalization ability and predictive power of the resultant classifier.

4.1.2 FOA algorithm

Fruit Fly Optimization Algorithm (FOA) is one of the global optimization methods that were developed on the basis of natural fruit fly foraging behavior. Through an artificial simulation of the way these insects find food sources based on sensory perception and successive position changes, the algorithm converts their biological behaviors into an organized search process capable of being applied computationally. The basic operational procedures of FOA are as below:

Inputs: maximum number of iterations K of fruit flies, population size N of fruit flies, search step R.

Step 1: Randomly select the initial position (X_0, Y_0) of the fruit fly population.

Step 2: With the first position (X_0, Y_0) as the starting point, all members of the fruit fly population will move independently to find food in a randomly selected direction and a randomly selected distance. The last position achieved by the i th fruit fly after such stochastic displacement is referred to as (X_i, Y_i).

$$\begin{cases} X_i = X_0 + 2 * R * rand(\cdot) - R \\ Y_i = Y_0 + 2 * R * rand(\cdot) - R \end{cases} \tag{18}$$

$rand(\cdot)$ stands for a random value that is randomly selected to lie between $[0,1]$.

Step 3: The distance between the current location of each fruit fly within the population and the origin is calculated. The next step is to take the reciprocal of this distance and use it as the flavor concentration judgment value P_i of that particular fruit fly:

To every individual in the population, the Euclidean distance between their initial position and the origin is calculated and the inverse of that distance is considered to be the flavor concentration measurement value AA (per fruit fly):

$$P_i = \frac{1}{\sqrt{X_i^2 + Y_i^2}} \tag{19}$$

Step 4: The flavor concentration S_i of every fruit fly is measured by putting its corresponding judgment value in place in the concentration determination function. The comparison is then done between all the individuals in the population to find out which fruit fly has the highest flavor concentration. The concentration value and the positional coordinates of the best individual are maintained in order to be used subsequently:

$$[S', I'] = \max(S_i) \tag{20}$$

Step 5: This maximum value of flavor concentration found in Step 4 is retained, and the other members of the fruit fly colony reorient themselves and move to the position linked to that highest concentration. The first positional reference is updated appropriately to (X_0, Y_0):

$$\begin{cases} X_0 = X_{I'} \\ Y_0 = Y_{I'} \end{cases} \tag{21}$$

Step 6: Step 2-4 are repeated. If the maximum flavor intensity found in the current iteration is higher than that of the previous generation and the total number of iterations has not exceeded a predetermined limit K, the algorithm moves on to Step 5 to keep searching. When no longer does one or both of these conditions hold, the algorithm stops and outputs the most favorable concentration of flavors that have been registered during the whole optimization procedure.

4.1.3 SVM parameter optimization based on improved FOA

The FOA algorithm is improved in the present research through the introduction of the idea of adaptive step size. With this change, the search step size can be adjusted dynamically depending on the ratio of the current optimal flavor concentration value and the maximum flavor concentration value of the previous generation.

When the number of iterations $j = 1$, let $R(1) = R(2) = 2$ and $S'(1)$ be the highest flavor concentration value found by the first iteration of the algorithm.

At the moment when the number of iterations achieves $j \geq 2$, the comparison is made between the maximum flavor concentration of the present iteration and the highest value measured in the last iteration. If the peak of the current iteration drops lower than the previous best, it means that the search has shifted to less favorable part. To counteract this, the step size is increased to increase the area of exploration and again enable the algorithm to perform global search. The step size $R(j+1)$ controlling the next iteration is:

$$R(j+1) = R(j)\left(1 + \frac{S'(j)}{S'(j-1)}\right) \tag{22}$$

On the other hand, if the peak flavor concentration of the current iteration is higher than the peak flavour concentration of the previous iteration, the search is going efficiently towards a better region. In these conditions, it is reasonable to decrease the step size since smaller positional changes would enhance the accuracy of the solution. The step size $R(j+1)$ used in the next iteration is then given by:

$$R(j+1) = R(j)\left(1 - \frac{S'(j-1)}{S'(j)}\right) \tag{23}$$

The enhanced FOA algorithm has been extended to simultaneously maximize the penalty coefficient γ and the Gaussian kernel parameter gamma in the Gaussian kernel SVM classifier framework. The process is organized in such a way.

Input: the highest possible number of iterations K of the fruit fly population, the population size N, an initial step size value $R(1)$ given at the first iteration, and the condition $R(1) = R(2)$.

Step 1: Since the Gaussian kernel SVM framework has two parameters that need to be optimized simultaneously, two separate fruit fly populations are used simultaneously. Population 1 starts at starting point (X_{01}, Y_{01}) and population 2 at starting point (X_{02}, Y_{02}) where the first position of each group is randomly chosen.

Step 2: Use random direction and distance to search for the location of food. Let the i th fruit fly in the first group of fruit fly population find a new location (X_{i1}, Y_{i1}), and the i th fruit fly in the second group of fruit fly population find a new location (X_{i2}, Y_{i2}), then:

$$\begin{cases} X_{i1} = X_{01} + 2 * R(j) * rand(\cdot) - R(j) \\ Y_{i1} = Y_{01} + 2 * R(j) * rand(\cdot) - R(j) \\ X_{i2} = X_{02} + 2 * R(j) * rand(\cdot) - R(j) \\ Y_{i2} = Y_{02} + 2 * R(j) * rand(\cdot) - R(j) \end{cases} \tag{24}$$

Within the formula, $rand(\cdot)$ is a randomly generated value selected inside the range $[0,1]$ and j stands for the number of current iterations.

Step 3: The distance to the origin is calculated on all the fruit flies of both the populations and the reciprocal of each distance is considered as the flavor concentration judgment value P of the respective individual. The calculation is done independently between each fruit fly in the first population and each fruit fly in the second population:

$$\begin{cases} P_{i1} = \dfrac{1}{\sqrt{X_{i1}^2 + Y_{i1}^2}} \\ P_{i2} = \dfrac{1}{\sqrt{X_{i2}^2 + Y_{i2}^2}} \end{cases} \tag{25}$$

Step 4: The results obtained through the optimization process are assigned to the penalty coefficient γ and the kernel parameter gamma in the Gaussian kernel SVM classification model.

$$\begin{cases} \gamma = 10 P_{i1} \\ gamma = P_{i2} \end{cases} \tag{26}$$

Step 5: The classification accuracy of the Gaussian kernel SVM model is used as the flavor concentration decision function. In every population group of fruit flies, the flavor concentration values are compared between all individuals to find the group with the highest

concentration of the present iteration. The concentration value $S'(j)$ and positional coordinates $I'(j)$ of the optimal group are stored to be used later. Formally:

$$[S'(j), I'(j)] = \max S_i \tag{27}$$

Step 6: If the number of iterations j is equal to 1, perform step 7; if the number of iterations j is greater than 1, continue to determine whether the best flavor concentration $S'(j)$ of this iteration is greater than $S'(j-1)$, and if it is greater than $S'(j-1)$ then make the next iteration step length $R(j+1) = R(j)(1-s'(j-1)/s'(j))$, and if it is less than equal to $S'(j-1)$ then make the next iteration step $R(j+1) = R(j)(1+s'(j)/s'(j-1))$.

Step 7: In case where the number of iterations j is 1, $S'' = S'(1)$ is directly assigned. If the number of iterations j is greater than 1, another conditional check is run to ascertain if $S'(j)$ is higher than S''. In such instance, $S'' = S'(j)$ is set, the value of γ' and $gamma'$ at that time is taken down, and all fruit fly populations are steered towards the location which corresponds to the greatest flavor concentration found during the present iteration. If DDD is however smaller than or equal to EEE, no change happens and the two groups move in the same direction towards the position representing the best flavor concentration of the current iteration. This process can be described as:

$$\begin{cases} X_{01} = X_{I_1'} \\ Y_{01} = Y_{I_1'} \\ X_{02} = X_{I_2'} \\ Y_{02} = Y_{I_2'} \end{cases} \tag{28}$$

Step 8: The check carried out at step number 8 is to see if the current iteration count j has attained or gone beyond the set maximum K. When the limit is exceeded, the algorithm stops. Otherwise, assuming that the count is still less than the maximum level K, the algorithm repeats steps 2-7 in the next iteration.

The output is the best penalty coefficient γ' and the best kernel parameter $gamma'$.

4.2 Analysis of the results of student achievement prediction

When the model construction and parameter optimization have been completed, the forecasted grade trends produced by the model are compared with the real grade records of 324 second-year students of the XX Vocational College during 2023-2025. The direct visual comparison of predicted and observed values is shown in Figure 9 and the magnitude of prediction error is illustrated in Figure 10 using the same sample. The outcomes are favorable in both respects. The fitted model follows the general tendency of the academic performance quite closely, and it does not show significant gaps between expected and real score differences within the range of observations. Such results prove that the SVM classifier optimized based on the enhanced FOA can produce credible and prompt alerts about the risks posed by students in the academic field, giving the suggested framework some credibility in practice.

In addition, through the student academic risk early warning system constructed in this paper, it can better integrate behavioral data such as student consumption, Internet hours and library borrowing, and can push analytical warning signals to college management departments, realizing the management transformation from post-processing to active intervention. Through

data-driven, it promotes the management of higher education from empirical judgment to precise governance, from unified control to personalized support, and builds a whole-process management system of monitoring-warning-intervention-optimization.

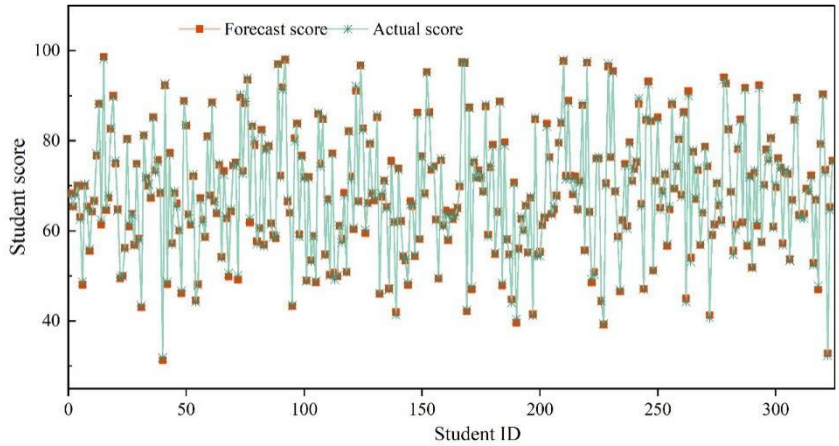

Figure 9: Comparison of predicted and actual scores

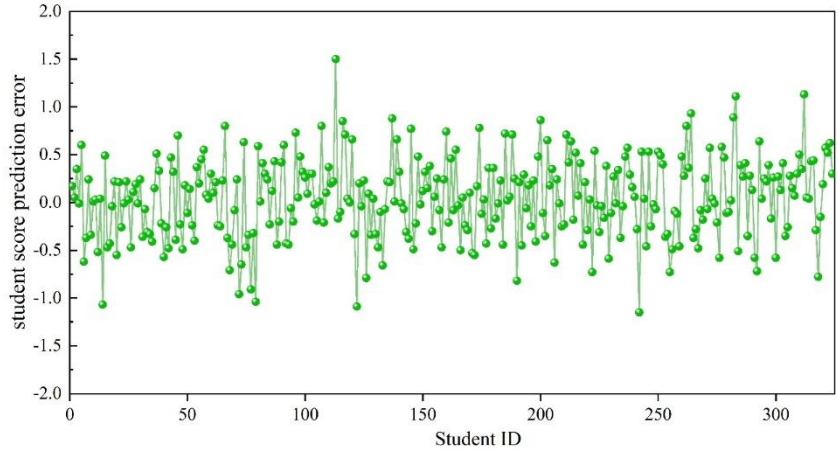

Figure 10: The error of comparing the predicted score with the actual score

5 Conclusion

The paper starts with the use of the improved K-means algorithm to group student behavior. Then, the changed Apriori algorithm is used to identify the correlation rules between student behavior and academic performance. Lastly, the support vector machine (SVM) parameters are used, which have been optimized with the help of an improved FOA, to enhance the early warning model in predicting the academic grades of students. The main findings of the research are:

In this paper, students' consumption at school, Internet access, bathing behavior in the bathhouse and borrowing of graph data were clustered according to three levels: low, medium and high, and it was found that 62.65% of the students had an average monthly consumption mean of 409.45 yuan, 43.52% of the students had a monthly consumption of 37.64 yuan for the campus Internet, and 48.77% of the students had an average monthly bathing in the range of 9-17 times, respectively, 73.15% of students borrowed an average of 6.67 books per month.

In terms of computational efficiency, the suggested Apriori variant is significantly better than any other benchmark algorithms when used to discover connections between student

actions and academic performances. Substantively speaking, the analysis has produced 37 unique association rules that connect behavior patterns to grade performance. The rules have an immediate practical use: college administrators or student affairs offices could rely on the typical profiles contained in each rule to develop specific plans of cultivation and individualized management models of students with the same behavioral signature.

The academic risk warning model proposed in this paper shows a significant ability to predict student performance trends depending on the observable learning actions. In addition to its predictive value, the framework has wider implications regarding the way vocational colleges approach student governance. This is an example of moving towards the notion of basing judgment on intuition to data-driven accuracy, and shifting away from one-size-fits-all administration to support systems that are truly sensitive to the needs of each student.

References

[1] McPherson, A., & Buchanan, R. (2019). Teachers and learners in a time of big data. Journal of philosophy in schools.

[2] Florea, D., & Florea, S. (2020). Big data and the ethical implications of data privacy in higher education research. Sustainability, 12(20), 8744.

[3] Veldkamp, B., Schildkamp, K., Keijsers, M., Visscher, A., & de Jong, T. (2021). Big data analytics in education: big challenges and big opportunities. International perspectives on school settings, education policy and digital strategies: A transatlantic discourse in education research, 266.

[4] Xing, W., & Wang, X. (2022). Understanding students' effective use of data in the age of big data in higher education. Behaviour & Information Technology, 41(12), 2560-2577.

[5] Yu, W. (2021, September). Application of Big Data Technology in the Innovation of University Education Management Work. In 2021 4th International Conference on Information Systems and Computer Aided Education (pp. 988-992).

[6] Liu, J. (2021, February). Research on the innovation of graduate educational administration management mode under the background of big data. In Journal of Physics: Conference Series (Vol. 1744, No. 3, p. 032026). IOP Publishing.

[7] Lv, Z. (2022). Research on optimization and application of university student development and management strategy driven by multidimensional big data. Scientific Programming, 2022(1), 6538069.

[8] Wu, S. (2023). Research on innovation and development of university instructional administration informatization in IoT and big data environment. Soft computing, 27(24), 19075-19094.

[9] Jin, J. (2025). Research and Construction of Student Management Platform for Special Needs Students with Decision Tree Model and Big Data Technology. Systems and Soft Computing, 200310.

[10] Peng, Y. (2023). Leveraging Big Data Technology for Enhanced University Education Management: A Path Analysis with Focus on Commercialization and Innovation

Strategies. Journal of Commercial Biotechnology, 28(2).

[11] Liang, Y., Wang, J., & Wang, J. (2025, June). Reconstructing Higher Education in the Big Data and AI Era: Interdisciplinary Integration and Problem-Driven Talent Cultivation. In 2025 International Conference on Distance Education and Learning (ICDEL) (pp. 47-51). IEEE.

[12] Wang, J. (2022). Optimization design of international talent training model based on big data system. Frontiers in Psychology, 13, 949611.

[13] Li, Y., & Wu, Y. (2024). Research on optimizing teaching strategies for financial and accounting talents based on big data analysis. Accounting, Auditing and Finance, 5(1), 37-44.

[14] Zeng, F., Xing, C., & Song, X. (2025). Optimizing educational dynamics: A big data approach to tailored teaching and enhanced student management. Journal of Computational Methods in Sciences and Engineering, 25(1), 514-525.

[15] Xin, X., Shu-Jiang, Y., Nan, P., ChenXu, D., & Dan, L. (2022). Review on A big data-based innovative knowledge teaching evaluation system in universities. Journal of innovation & knowledge, 7(3), 100197.

[16] Meng, Q., Sun, C., & Li, D. (2023, November). Monitoring and Evaluation of Talent Cultivation Quality Based on Big Data. In International Conference on Cognitive based Information Processing and Applications (pp. 509-521). Singapore: Springer Nature Singapore.

[17] Chen, Y., Zheng, W., Huang, X., Liu, C., & Lin, X. (2025). Big Data and AI-Driven Outcome-Based Education Talent Cultivation in General Universities: Developing a Dual-Dimensional “Process-Outcome” Evaluation System. Journal of Sociology and Education, 1(8).