International Journal of Gaming and Computer-Mediated Simulations
Volume 16 • Issue 1 • January-December 2024

Simulation-Based Machine Learning for Predicting Academic Performance Using Big Data

Cheng Zhang
Shaanxi University of Science and Technology, China

Jinming Yang
Shaanxi University of Science and Technology, China

Mingxuan Li
Shaanxi University of Science and Technology, China

Meng Deng
Shaanxi University of Science and Technology, China

ABSTRACT

In this study, simulation and big data analytics are combined with machine learning techniques, specifically K-means clustering, Apriori algorithm, and a stacked integrated learning model, to predict academic performance of college students with a high accuracy of 95.5%. By analyzing behavioral data from over 1,000 undergraduates, we correlate various behaviors with academic success, focusing on the use of libraries, self-study habits, and internet usage. Our findings highlight the benefits of using big data and simulation in educational strategies, promoting effective resource allocation and teaching enhancements. The study acknowledges limitations due to its regional focus and proposes future research directions to enhance model generalization and technological integration for broader application.

KEYWORDS

Academic Performance, Apriori Algorithm, Big Data, Integrated Learning, K-Means Clustering, Machine Learning, Simulation

INTRODUCTION

In the context of the information age, all industries have accumulated a vast amount of data. This data often implies valuable knowledge and information. Through machine learning and data mining technologies, we can uncover inherent rules and patterns within the data from which to extract valuable insights and knowledge (Zhen, 2022). Thus, they can solve problems in various fields and help managers to make decisions in a more "scientific" way. Effective use of data and intelligent, rational analysis of data should become the consensus of the academic community (Ma, 2020). At present, machine learning and data mining have been widely used in education, e-commerce, intelligent manufacturing, transport, and other fields.

Similarly, student management and training in colleges and universities need not only empirical guidance but also scientific guidance. If college administrators can use data effectively to make scientific decisions, the quality of management work will be substantially improved (Wang et al., 2024).

DOI: 10.4018/IJGCMS.348052

In this context, educational data mining came into being. Its purpose is based on the massive data resources accumulated in the field of education, the use of educational psychology, computer science and statistics, and other disciplines of theory and technology to find and solve a variety of problems in educational research and teaching practice (Sun, 2023). In recent years, due to the widespread use of Internet applications, the eating, drinking, living, and travelling of students in colleges and universities are inseparable from the use of the Internet, which means that students will produce a large number of behavioral data every day. Beneficial analysis and mining of these behavioral data can help university leaders make more informed decisions, optimize resource allocation, improve the quality of education, promote the overall development of students, provide a new direction for the study of student management, and have a far-reaching impact on the educational management of higher education institutions. For example, the on-campus network generates a large amount of data every day, including network data, consumption data, attendance data, academic performance, book borrowing data, and more. These data reflect rich information about students' thinking, emotions, and behavioral dynamics. How to store and manage these massive data scientifically and analyze them effectively is of great significance for the development of colleges and universities and the future of students. Student performance prediction will be one of the most important research branches of educational data mining (Ma, 2020; Du, 2023; Pan, 2020; Liu, 2020).

Student performance prediction, also known as student academic performance prediction, is one of the important research problems in the field of educational data mining. It aims to predict the future academic performance of students by using student-related information, such as course grades, grade point average (GPA), risk of failing a course, and risk of dropping out. The use of big data for analysis and mining technology, as well as scientific analysis and prediction of education big data, is a good way to ensure the smooth implementation of teaching and improve the effectiveness of teaching and learning, while also developing future educational trends. If we use students' learning behavior data, relying on machine learning methods to establish performance prediction models, the established performance prediction models can support teachers and students to adjust teaching strategies dynamically and effectively optimize education and teaching. Learning achievement prediction can not only help teachers amend teaching strategies in time, develop or improve teaching strategies, improve teaching methods, and optimize the allocation of resources so as to improve the output of education and teaching to improve students' final academic performance and reduce the proportion of failing students, but also to a certain extent play a supervisory and early warning role for the students; it is therefore an effective method of improving students' performance (Cao, 2022). Furthermore, for students with the risk of failing a course or repeating a grade or dropping out of school, the research on student performance prediction is particularly significant. Teachers or administrators can take appropriate measures based on the prediction results, such as giving extra attention and guidance to “at-risk students,” so that these students can avoid course failure or even ultimate academic failure.

Therefore, the performance prediction model uses a large amount of educational data and applies technologies such as data mining and machine learning to analyze and mine students' learning characteristics, behaviors, emotions, and other factors, as well as the difficulty and relevance of the course, so as to build an effective prediction model predict students' future learning performance. Advancements in predictive analytics have significantly influenced educational strategies, with a notable shift towards data-driven methodologies for assessing academic performance. Zeineddine et al. (2021) and Abdul Bujang et al. (2021) have highlighted the use of sophisticated machine learning techniques, including Automated Machine Learning and various traditional models, to predict student outcomes with greater accuracy. These studies provide valuable insights into the effectiveness of modern educational technologies and set the stage for further exploration in our research. This can help teachers understand the learning situation of each student, provide personalized teaching support and intervention, and improve the quality and effectiveness of teaching. At the same time, it can also help students to self-assess and adjust their learning strategies, increase their interest and motivation in learning, and promote their academic performance (Li, 2023; Zhai, 2022; Xu, 2023; Ren & Yang, 2020; Su, 2019).

In this paper, in agreement with most researchers and based on the analysis of a large amount of behavioral data exploring students' basic attributes, behavioral habits, learning attitude representations, and more, we classify students' behavioral characteristics on campus through the K-means algorithm in cluster analysis. We then use the Apriori algorithm to correlate students' behavioral characteristics with their academic performance; the students' behavioral prediction model based on behavioral characteristics can effectively predict students' academic performance. This means that we can finally predict future performance of college students empowered by big data through mathematical and scientific calculations to achieve college student management in the context of big data, which is of great significance when improving the quality of education, optimizing teaching resources, promoting student development, and realizing personalized education.

RELATED WORK

Recent studies have expanded the applications of big data across various fields. Meng et al. (2023) analyze the influence of Internet celebrity propaganda on purchasing decisions using big data, highlighting its role in consumer behavior analysis. Wang et al. (2023) and Zhang et al. (2023) explore big data's utility in enhancing enterprise information security management and optimizing supply chain performance, respectively, underscoring its effectiveness in risk assessment and logistical operations. Additionally, Gao et al. (2023) employ social media big data for global market demand forecasting, demonstrating its potential to refine marketing strategies and product development. These works collectively illustrate the transformative impact of big data in driving strategic decision-making and operational efficiencies across sectors.

Academics have carried out early and rich research related to college performance prediction, which has laid a solid foundation for subsequent in-depth research. As early as the beginning of the 21st century, foreign scholars conducted in-depth research on student achievement prediction from the perspective of educational data mining and machine learning. Especially in the last decade, the depth and breadth of the research on this issue has become more and more detailed. Pardos et al. (2010) proposed a knowledge tracking model based on Bayesian network, which uses the question-answering data of students in the online education platform to model and predict their knowledge mastery and makes personalized teaching recommendations to students based on the prediction results. Li and Fu (2010) systematically analyzed and summarized the development history, research contents, application fields, technical methods, and development trends of educational data mining at home and abroad, and they proposed the application prospects and problems of educational data mining in China's education field. Thai-Nghe et al. (2011) proposed a "matrix decomposition based model," which uses a matrix decomposition technique to decompose the student-course rating matrix into two low-rank matrices representing the implicit features of the student and the course, respectively, and then uses these features to predict the student's grades in other courses. Huang (2012) used the likelihood weighting algorithm in Bayesian nets to predict the student's course assessment grades based on the student's previous course grades. Some other researchers have conducted studies using rule-based generation methods. Huang and Fang (2013) used a decision tree algorithm to predict the assessment grade on course "dynamics" based on the students' previous comprehensive GPA and the grades of the four precursor courses. Piech et al. (2015) proposed a "deep knowledge tracking model" using a deep neural network model to track students' mastery of different knowledge points and predict their performance in the future based on their answer records on an online education platform. Meier et al. (2016) used homework completion and midterm exam results and course project completion to predict students' final exam performance. Ren et al. (2016) built a prediction model for students' performance based on "length of video viewing," "average number of modules studied per day," and "number of quizzes completed" on the catechism platform. Minn et al. (2018)

proposed a model based on multi-task learning and graph neural networks to predict students' mastery on different knowledge points simultaneously, and they used graph neural network method to model the dependency relationship between knowledge points, which improved the understanding and prediction of students' knowledge status. Feng et al. (2019) clustered students based on their learning behaviors and used a convolutional neural network to fuse students' individual learning behaviors, the learning behaviors of others in the same category, and course information to predict whether students can complete the course. Yao et al. (2021) used a BP neural network to model and analyze the factors affecting students' grades in colleges and universities and to predict students' final grades. Hu and Zhao (2021) used a decision tree algorithm to select, classify, and preprocess the raw data of a foreign high school grade, and they implemented the decision tree algorithm by using software such as SKlearn to generate offline models and academic indicator assessment reports; they finally predicted the students' grades. Hidalgo et al. (2021) explored the potential of deep learning and meta-learning to predict the potential of student performance in virtual learning environments. They implemented a prediction model that automatically optimizes the architecture and hyperparameters of a deep neural network and conducted experiments on a dataset containing more than 500 online university master's students. The results show that the model performs comparably to traditionally designed models, with greater efficiency and scalability.

In summary, although achievement prediction research has made great progress, there are still some problems. Firstly, there is the problem of data quality, which is an important factor affecting the accuracy of achievement prediction models. Then there is the problem of feature selection. Feature selection refers to selecting a subset of features with high correlation with the target variables from daily behavioral data, increasing the behavioral data features related to performance prediction and improving the generalization ability of the model, which needs to be reasonably selected and optimized according to different data characteristics and model requirements. Finally, there is the model selection problem. Model selection refers to choosing an optimal model from multiple candidate models to achieve the best prediction effect.

MATERIALS AND METHODS

In addressing the challenge of predicting academic performance, our study introduces a novel simulation-based machine learning approach. This methodology involves generating synthetic datasets which accurately reflect the diversity and complexity of student behaviors. By doing so, we enhance the training process of our machine learning models, ensuring that they are not only accurate but also capable of generalizing across different student populations and scenarios. Our simulation-based approach involves the creation of synthetic data which simulates a wide range of possible academic outcomes based on observed student behaviors. This data is then used to train a suite of machine learning models, including Random Forest, Gradient Boosting Trees, and XGBoost. Each model learns to predict academic performance based on a comprehensive set of features derived from both real and synthetic datasets. The integration of synthetic data helps in testing the models against diverse scenarios, significantly enhancing their predictive accuracy and robustness.

Data Preprocessing

Data Sources and Interpretation

In this study, we conducted detailed data preprocessing and analysis of the behavioral data of more than 1,000 undergraduate students at S University in western China. Our aim was not only to explore the correlation between students' behavioral data and academic performance but also to ensure the fairness and unbiased nature of our predictive models. To this end, we implemented several key strategies during data preprocessing: Firstly, we ensured that our dataset accurately represented the demographic diversity of the student population, incorporating balanced sampling techniques.

Table 1. Corresponding values of student behavior characteristics

No.	Norm	Corresponding value				
A	Average monthly frequency of library visits	1. Less than 5	2. 5 to 10	3. 11 to 15	4. 16 to 20	5. 20 or more
B	Average number of times borrowed per month	1. 0 times	2. 1 time	3. 2 to 3 times	4. 4 to 5 times	5. More than 5 times)
C	Average monthly hours of library access	1. Less than 5 hours	2. 5 to 10 hours	3. 11 to 20 hours	4. 21 to 30 hours	5. More than 30 hours
D	Class attendance	1. Less than 60%	2. 60-70%	3. 71-80%	4. 81-90%	5. 91-100%
E	Self-study attendance	1. Less than 60%	2. 60-70%	3. 71-80%	4. 81-90%	5. 91-100%
F	Homework completion	1. Less than 60%	2. 60-70%	3. 71-80%	4. 81-90%	5. 91-100%
G	Activity participation rate	1. Less than 60%	2. 60-70%	3. 71-80%	4. 81-90%	5. 91-100%
H	Activity check-in rate	1. Less than 60%	2. 60-70%	3. 71-80%	4. 81-90%	5. 91-100%
I	Activity check-out rate	1. Less than 60%	2. 60-70%	3. 71-80%	4. 81-90%	5. 91-100%
J	Activity tardiness rate	1. Greater than 60%	2. 40-60%	3. 21-40%	4. 11-20%	5. Less than 10%
K	Weekly frequency of getting up early for breakfast	1. 0 times	2. 1 to 2 times	3. 3 to 4 times	4. 5 to 6 times	5. 7 times
L	Average monthly frequency of getting up earlier than 8 o'clock	1. Less than 5 times	2. 5 to 10 times	3. 11 to 15 times	4. 16 to 20 times	5. More than 20 times
M	Average monthly time spent on the internet	1. Less than 5 hours	2. 5 to 10 hours	3. 11 to 20 hours	4. 21 to 30 hours	5. More than 30 hours
N	Average monthly expenditure	1. Less than $800	2. $801-1200	3. $1,201-1,600	4. $1,601-2,000	5. More than $2,000
O	Parents' highest educational attainment	1. Junior high school and below	2. High school/ secondary school	3. Junior college	4. Undergraduate	5. Postgraduate
P	Annual household income status	1. Less than 50,000	2. 5 to 100,000	3. 100 to 200,000	4. 200 to 400,000	5. 400,000 or more

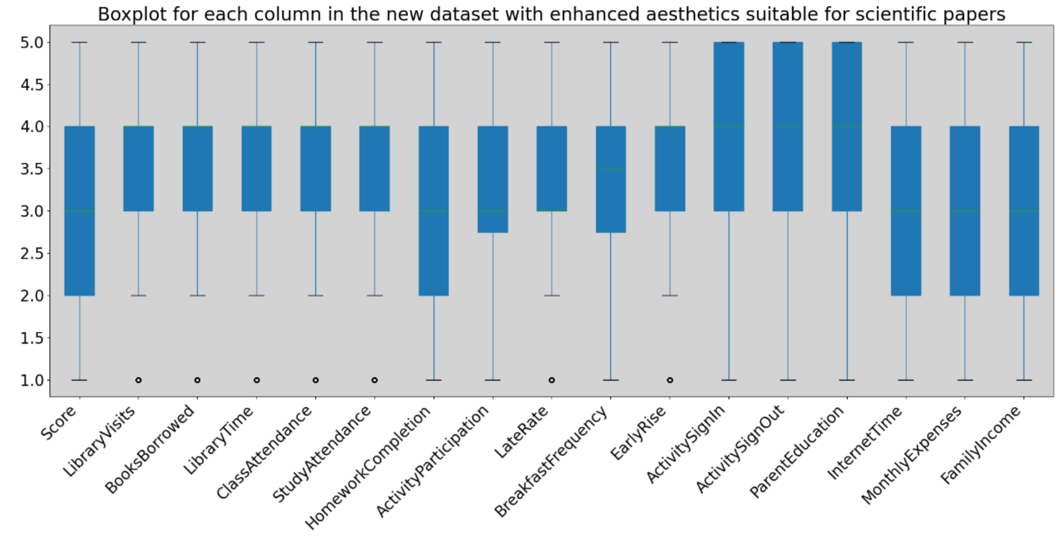
Secondly, we conducted rigorous fairness testing during the model validation phase to identify and mitigate any potential biases that could affect various student groups. This involved analyzing prediction outcomes by different demographic subgroups to ensure that no group was unfairly advantaged or disadvantaged. These steps are crucial for applying big data in a way that promotes fair and accurate student management.

The dataset was thoroughly checked for completeness and accuracy. Each field in the dataset was meticulously examined to determine the type and format of the data. The majority of fields

Table 2. The corresponding value of students' GPA

Norm	Corresponding value				
Student's grade point average (GPA)	1. Less than or equal to 1.00	2. 1.01-2.00	3. 2.01-3.00	4. 3.01-3.99	5. 4.00 and above

Figure 1. Dataset box-whisker plot



were identified as integer types, indicating that the data had been appropriately coded to facilitate quantitative analysis.

After confirming the format and structure of the data, further outlier detection was performed. By applying statistical methods such as box-and-line graph analysis, we visualized and analyzed the distribution of each field in the dataset. This step helped us to identify possible outliers in the data, thus providing important reference information for further analysis of the data. The Box-whisker Plot is shown in Figure 1.

K-means Clustering Algorithm

In our study, which aims to predict students' performance intervals from their behavioral characteristics, the K-means clustering algorithm provides a strong methodological basis. K-means is a widely used non-hierarchical clustering algorithm, and we use the K-means algorithm to group the students based on their behavioral characteristics and divide the data into K disjoint subsets (clusters). Each subset has a high degree of similarity, while the data between different subsets has a low degree of similarity. By identifying and analyzing groups of students with similar behavioral characteristics, we can better understand which behavioral habits may be associated with high or low grades. In practice, the K-means algorithm helps us identify groups of students with similar behavioral patterns by minimizing the sum of the distances between the data points within each cluster and the cluster center, which is a representative of the group's behavioral characteristics. The algorithm first randomly selects K initial cluster centers and then, through an iterative process, continuously adjusts the attribution of these centers and data points to ensure that each student is assigned to the group that is closest to their behavioral characteristics. This process is based not only on students' behavioral characteristics such as frequency of library use, number of borrowings, or class and self-study attendance, but also a combination of their lifestyle habits, such as frequency of breakfasts and number of early mornings, as well as contextual factors, such as family financial status and parental qualifications.

K-means tries to minimize the total internal variance of all the clusters. That is,

$$\mathrm{J} = \sum_{k=1}^{K} \sum_{i \in C_k} x_i - \mu_k{}^2 \tag{1}$$

where J is the objective function and our goal is to minimize J, C_k is the set of data points in the kth cluster, x_i is the data points belonging to C_k, and μ_k is the centre of C_k and is calculated as the average of all the points within the cluster.

The algorithm steps are as follows:

1. Initialization: randomly select K data points as initial clustering centers.
2. Assignment step: for each data point, assign it to the cluster to which the nearest cluster center belongs according to the minimum distance criterion.
3. Update step: recalculate the center of each cluster, that is, the mean of all data points in each cluster.
4. Iteration: repeat the assignment and update steps until the clusters no longer change or a predefined number of iterations is reached.

The mathematical explanation is as follows:

1. Distance Measure:$x_i - \mu_k{}^2$ is the square of the Euclidean distance used to measure the distance from data point x_i to the cluster centre μ_k. By minimising the sum of this distance, the algorithm tries to keep the data points within a cluster as close to the cluster center as possible.
2. Cluster center update: The update formula for the cluster centre μ_k ensures that the new cluster centre is the geometric centre of all points within the current cluster, which helps to reduce the variance within the cluster.
3. Convergence: the K-means algorithm gradually optimizes the objective function J through an iterative process until it reaches a steady state. Theoretically, this process ensures that the algorithm will converge to a local minimum, but the final result may depend on the choice of the initial cluster centres.

Through this refined clustering, the K-means algorithm gives us insight into the statistical links between specific behavioral habits and student performance intervals. For example, the algorithm helps us identify which behavioral traits are strongly associated with excellent grades (4.00 and above) and which are likely to lead to poor grades (less than or equal to 1.00). This insight is extremely valuable to educators because it provides possible ways to intervene and promote student academic success.

Apriori Algorithm

Applying the Apriori algorithm on preprocessed data, our first task is to identify frequent itemsets. These frequent itemsets represent patterns which are common in student behavior data. A frequent itemset is a set of items that occur more frequently (or with more support) than a user-specified minimum support threshold across all transactions. This means that if we have a large database of transactions, the frequent itemset is the set of those items that frequently occur together in this database.

The algorithm is based on two key metrics:

- support, a measure of how often the itemset occurs in the entire dataset, where

$$\text{Support(X)} = \frac{\textit{Number of transaction containing X}}{\textit{Total number of transactions}} \tag{2}$$

- confidence, a measure of the accuracy of a rule, where

$$\text{Confidence(X} \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} \tag{3}$$

where X and Y are sets of terms and X ∩ Y = ∅.

Algorithm Steps

In the Apriori algorithm, minimum support and confidence are key parameters. Support helps us determine which itemsets are "frequent" enough, while confidence measures the strength of the association between these itemsets and a high or low GPA. In our study, it is crucial to set these thresholds appropriately because they determine the quality and usefulness of the final rule. Too low a threshold may result in too many irrelevant rules, while too high a threshold may miss valuable patterns. After identifying the frequent itemsets, we will use the Apriori algorithm to generate association rules. These rules will reveal how different behavioral patterns are associated with academic performance. For example, a possible rule would be: "If a student has high class attendance and high study hall attendance, they have a high probability of earning a GPA of 4.0 or higher." These rules not only help us to understand which combinations of behaviors are positive but may also reveal patterns of behavior that need to be improved. After generating the rules, we need to evaluate them to ensure their validity and usefulness. K-means clustering and Apriori algorithm were chosen for their ability to handle large-scale data and identify patterns and relationships within the data effectively. K-means provided a preliminary grouping of data based on student behaviors, which facilitated targeted analysis using the Apriori algorithm to uncover association rules between behaviors and academic performance.

A Behavioral Trait-Driven Model for Predicting Student Academic Achievement Based on Stacked Integrated Learning

Classification Of Student Behavioral Characteristics By K-means Algorithm

In this study, we used the K-means clustering algorithm to analyze and classify the behavioral characteristics of students on campus. The K-means algorithm is a widely used clustering technique which effectively divides the data into clusters so that data points within the same cluster are as similar as possible and data points within different clusters are as different as possible.

1. Data preprocessing: Firstly, we normalized the dataset. K-means algorithm is very sensitive to the scale of the data when calculating the distance between data points, so we normalized all numerical features to ensure that each feature has the same weight in the clustering process.
2. Determining the number of clusters (K): In order to determine the optimal number of clusters, we use the elbow rule. By calculating the sum of squares (SSE) for different values of K, we look for the *elbow point* of the SSE curve, that is, the inflection point where the SSE starts to drop sharply, as a reference for the optimal number of clusters.
3. Cluster analysis: After determining the optimal number of clusters, the K-means algorithm was applied to cluster the student behavior data. Each student was assigned to the cluster represented by its closest centroid. Once this step was completed, it was possible to observe differences in student behavioral characteristics across clusters (or groups).
4. Analysis of clustering results: For each cluster, we calculated the mean value containing each characteristic to depict the profile of behavioral characteristics of each group. This information

helped us to understand the characteristics of the different groups of students and provided a basis for further analyses.

Associating Student Behavioral Characteristics with Academic Achievement Via Apriori Algorithm

After clustering the student behavioral traits, we further used the Apriori algorithm to explore the association between these behavioral traits and students' academic performance.

1. Data preparation: The raw student behavioral data and academic performance were converted into a format that could be efficiently processed by the Apriori algorithm. Each student's data record is considered as a *transaction* containing *items* representing different behavioral characteristics and achievement levels of the student. To do this, we first need to convert each record in the dataset into a list of itemsets.
2. Assume that the original dataset contains several columns containing students' behavioral characteristics such as age, gender, participation in activities, study habits, and their previous grades and expected grades. Each feature can be considered as an item, while each student record is a transaction containing multiple items.
3. Frequent itemset mining: Using the Apriori algorithm, we mine frequent itemsets from the transaction dataset. We set a minimum support threshold so that only frequent itemsets in the dataset are selected.
4. Generate association rules: For the mined frequent itemsets, we further generate association rules. These rules reveal potential associations between student behavioral traits and academic performance. We filter out the more meaningful rules by setting a confidence threshold.

These models aid teachers by providing insights into which student behaviors align with better academic results, allowing for targeted interventions and personalized teaching strategies that address students' specific needs and habits.

Construction Of Stacked Integration Models

Based on the results of K-means algorithm and Apriori algorithm, then next step is to construct the stacked integration model. A stacked integration model is a hierarchical model that is based on several different base learners whose predictions are used as inputs to another model (called a meta-learner).

The first step is selection and training of base learners. We have selected three powerful regression models as base learners. In constructing the predictive model for this study, the choice of machine learning algorithms was pivotal to address the complexities inherent in student behavioral data. Our decision to employ a stacked integration model incorporating Random Forest, Gradient Boosting Tree, and XGBoost was grounded in a strategic evaluation of each algorithm's strengths when handling diverse data challenges. Random Forest was selected for its robust performance across various data distributions, effectively handling the nonlinear relationships and complex interactions amongst the predictors without overfitting due to its ensemble approach which averages multiple decision trees trained on different parts of the data. Gradient Boosting Tree was included for its proficiency in optimizing previous trees' errors, progressively enhancing the model's accuracy. XGBoost added a layer of sophistication with its advanced regularization features, which mitigate overfitting. This is crucial given the multifaceted nature of behavioral data. Together, these algorithms complement each other, and each addresses specific shortcomings of the others, thereby fortifying the model's predictive capability. The meta-learner, implemented via a linear regression model, integrates the individual predictions from each base learner, yielding a harmonized final prediction. This approach not only leverages the distinct advantages of each algorithm but also enhances the generalizability and interpretability of the model, ensuring robust application across varied educational settings.

A Random Forest Regressor is an integrated learning method which makes final predictions by constructing multiple decision trees and outputting the average of their predictions. The mathematical formula for the Random Forest model is:

$$f_{RF}(X) = \frac{1}{B}\sum_{b=1}^{B} T_b(x;\theta_b) \tag{4}$$

where B is the number of trees, T_b is the prediction function for the bth tree, θ_b is the tree parameters and X is the input feature vector.

A Gradient Boosting Regressor is a boosting method which improves the accuracy of the model by iteratively training the decision tree to correct the residuals of the previous tree. The mathematical formula for the Gradient Boosting Tree model is:

$$f_{GB}(X) = \sum_{b=1}^{B} \eta_b T_b(x;\theta_b) \tag{5}$$

where η_b is the learning rate of the bth tree, and T_b andθ_b are defined as above.

XGBoost (eXtreme Gradient Boosting) is an efficient and flexible gradient boosting framework. Its core lies in the sequential construction of decision trees, where each tree learns and corrects the errors of the previous tree. XGBoost introduces regularization terms (L1 and L2 regularization) to reduce the complexity of the model and the risk of overfitting, and the mathematical formula can be expressed as:

$$\mathrm{Obj}(\theta) = \sum_{i=1}^{n} l(y_i,\hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

where l is the loss function which represents the difference between the model's predicted value $\hat{}$ y_i and the actual value y_i; $\Omega(f_k)$ is the regularization term, which is used to measure the complexity of the model, where f_k is the kth tree in the model; and K is the total number of trees in the model. In this way, XGBoost effectively controls the model complexity and improves the generalization ability while maintaining the accuracy of model prediction.

Next it is necessary to select and train the meta-learner. The predictions provided by the base learner are considered as new features which are used to train the meta-learner. In our model, we have chosen Linear Regression as the meta-learner.

Linear regression is a simple but effective algorithm which assumes a linear relationship between the features and the predicted results and finds the optimal coefficients by minimizing the mean square error between the predicted and actual values. The mathematical formula for a linear regression model is:

$$f_{LR} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m \tag{7}$$

where β_0 is the intercept term, $\beta_1,\ldots,\beta_m$ are the model coefficients, and $X_1,\ldots,X_m$ are the feature sets consisting of the predictions of the base learner.

In the stacked model, we merge the predictions of the base learners and use the meta-learner to make the final prediction of the merged predictions. The prediction of the entire stacked model can be represented by the following mathematical formula:

$$\hat{y}(X) = f_{LR}\left(\left[f_{RF}(X), f_{GB}(X)\right]\right) \tag{8}$$

where $\hat{y}(X)$ is the model's final prediction of student performance.

After completing the training of the model, we use a test set to evaluate the performance of the stacked model. The accuracy and explanatory power of the model predictions were quantified by calculating the mean square error (MSE) and the coefficient of determination (R^2). The formulas for MSE and R^2 are given below:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{9}$$

where n is the number of samples in the test set, y_i is the actual performance of sample i, and $\hat{y}_i$ is the model's predicted performance for sample i;

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{10}$$

where $\bar{y}$ is the average of all actual grades. the closer the R^2 value is to 1, the higher the prediction accuracy of the model.

These evaluation metrics allow us to understand the performance of the stacked model quantitatively in predicting students' academic performance, and then adjust and optimize the model to achieve better predictions.

In our study, different loss functions were chosen for each machine learning model within the stacked integration framework to optimize each model's performance according to its specific characteristics and learning algorithms. Random Forest, Gradient Boosting, and XGBoost each have unique approaches to processing data and learning from it, which means that they respond differently to various loss functions. Using the most suitable loss function for each model allows us to minimize errors more effectively and improve the robustness of predictions. For instance, XGBoost was paired with regularized gradient boosting to manage overfitting and variance, enhancing the generalization capabilities of our model. Utilizing different loss functions can provide a more tailored approach to training each model, which often results in better performance compared to using the same loss function across different types of models.

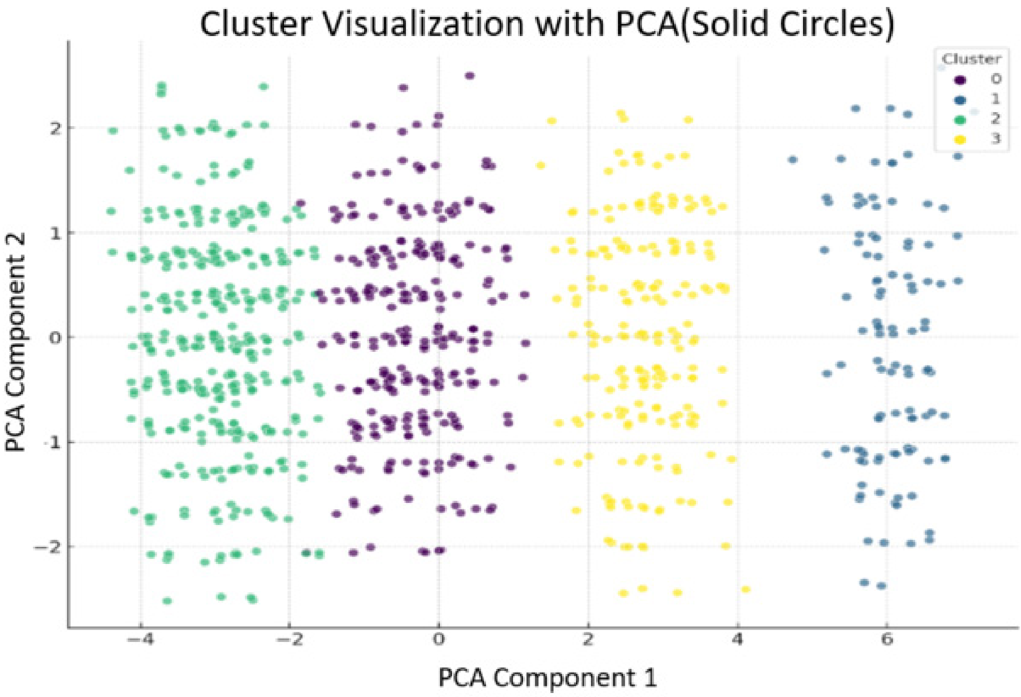
The K-means clustering and Apriori association rule mining introduced during the construction of the model provide additional insights into the model, enabling us to understand the factors affecting student performance in multiple dimensions.

The stacked integrated model incorporates the strengths of algorithms such as Random Forest, Gradient Boosting Tree, and XGBoost and combines the results analyzed by K-means clustering and Apriori algorithms in order to improve the accuracy of predicting students' expected grades. In the model design, the introduction of Random Forest increases the model's ability to deal with non-linear relationships; the sensitivity of Gradient Boosting Tree to details helps us to capture finer variations in the data; and XGBoost controls overfitting through regularization and ensures the model's ability to generalize.

RESULTS

In this study, we employed the K-means clustering algorithm and the Apriori algorithm in combination with various machine learning models such as Random Forest, Gradient Boosting Tree, and XGBoost to construct a stacked integration model to predict students' performance using their behavioral characteristics. First, we performed a comprehensive preprocessing of the dataset, followed by an effective classification of students' behavioral features by the K-means algorithm which divides students into different groups, each with a unique combination of behavioral features. Using the Apriori

Figure 2. Cluster of student behavioural characteristics



algorithm, we further analyzed the relationship between students' behavioral characteristics and their expected grades and mined out a series of meaningful association rules. All the processed data were subjected to model training. Eventually, the results were analyzed to verify the superiority of the prediction model and provide a model basis for subsequent research by universities and educators.

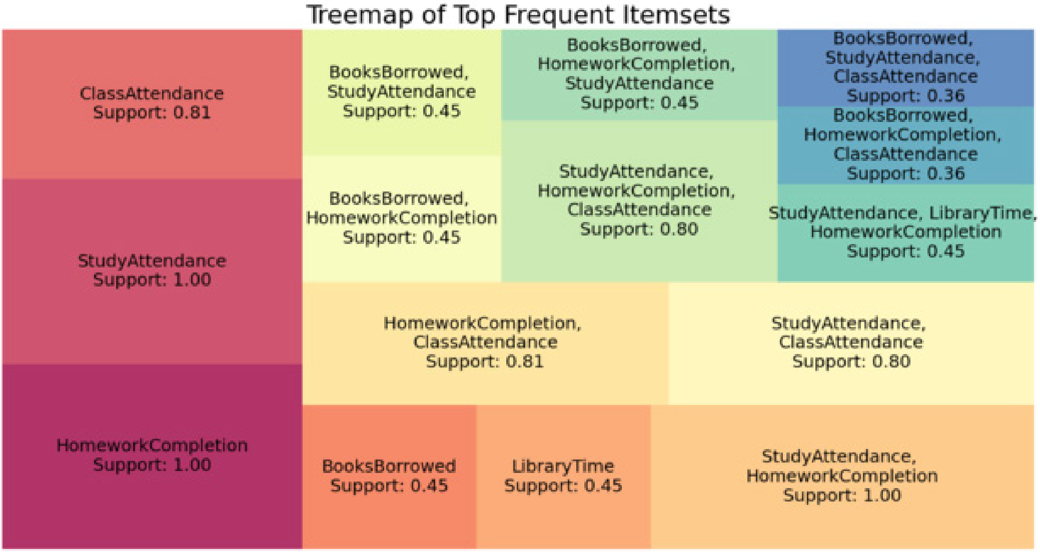
K-means Clustering and Apriori Association Rule Mining Analysis

We first used the K-means clustering algorithm to analyze and classify students' behavioral characteristics on campus. By normalizing the dataset and performing Principal Component Analysis (PCA), we effectively divided the student behavior data into four distinct clusters, as shown in Figure 2. Each cluster represents a group of students with a unique combination of behavioral characteristics.

Among the clustering characteristics are the following. Students in Cluster 0 show moderate academic achievement and balanced behavioral profiles, and they perform well in library activities and classroom attendance, but they are slightly underperforming in self-directed learning activities, implying that they perform better in a structured learning environment. Comparatively, students in Cluster 1 performed lower on all dimensions, particularly in academic achievement, suggesting that they face significant learning challenges and require additional support and intervention. Students in Cluster 2 performed well on most characteristics, particularly in terms of engagement and motivation, indicating high levels of motivation and self-driven learning. Finally, students in Cluster 3 performed moderately well on most learning behavioral characteristics, pointing to the fact that they needed additional attention and support in some areas.

In our study, we used the Apriori algorithm to mine the association rules between students' behavioral characteristics and grades. By analyzing the set of items associated with achievement grades of 4 and 5, we revealed potential links between students' behavioral patterns and achievement

Figure 3. Frequent itemsets tree diagram



expectations. As shown in Figure 3, the distribution of the number of frequent itemsets with different levels of support reflects the general trend of these association rules.

Specifically, in terms of individual characteristics, self-study attendance and homework completion showed the highest level of support (1.00), implying that these two behaviors are extremely common among the high-achieving student population (achievement levels 4 and 5). Similarly, the higher level of support (0.81) for class attendance again highlights the significant correlation between regular attendance and high student achievement. In terms of bimodal combinations of characteristics, the combinations of study attendance and book borrowing (self-study attendance, average number of times borrowed per month) and homework completion and classroom attendance (homework completion, class attendance) showed relatively high levels of support (0.45 and 0.81, respectively), which suggests that there may be a mutually reinforcing synergy between students' academic engagement and learning behaviors which together drive academic achievement. The combinations of study attendance and homework completion and class attendance and homework completion all demonstrated perfect support (1.00), further confirming the prevalence of these behavioral traits among high-achieving students. As for the combination of three characteristics, such as the combination of study attendance, library time, and homework completion (self-study attendance, average monthly hours of library access, homework completion), the level of support demonstrated was relatively low (0.45), but it was still the first among all the three combinations, suggesting that the linkage between these behavioral characteristics may have a positive impact on academic performance. Such findings help educators to deepen their understanding of the links between students' behavioral patterns and academic achievement, and they provide valuable guidance for promoting the integrated development of students' behavioral patterns.

Analysis Of Stacked Integration Model Results

We first used 80% of the data as a training set and 20% of the data as a test set, using Random Forest, Gradient Boosting Tree, and XGBoost as base learners. The loss functions are Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), which are used to train the Random Forest, Gradient Boosting Tree, and XGBoost models and continuously adjust

Table 3. Achievement prediction performance of stacked integration model

Level of achievement	Precision	Recall	F1-score	Support	Note
1	1.00	1.00	1.00	27	
2	1.00	0.97	0.99	38	
3	0.97	1.00	0.98	63	
4	0.95	0.91	0.93	57	
5	0.75	0.80	0.77	15	
Synthesis					Accuracy: 0.955
Macro average	0.93	0.94	0.93		
Weighted average	0.96	0.95	0.96		
Micro-average ROC AUC					0.971875

the parameters. The meta-learner uses linear regression and combines the predictions from the base learner to form the final prediction. The trained model is eventually saved and used to predict student performance.

In this study, an achievement prediction model based on student behavioral characteristics was developed. The model was optimized for prediction accuracy by constructing a soft voting classifier through a combination of Random Forest, Gradient Boosting Tree, and XGBoost algorithms. After optimizing the parameters by a grid search method, the final model achieved an overall accuracy of 95.5% on the test set, showing a high prediction performance. Table 3 shows the achievement prediction performance of the stacked integration model.

Achievement level indicates the level of student achievement predicted by the model. Overall accuracy indicates the overall accuracy of the model on all test data. Macro average indicates the simple average of the metrics on each achievement level. Weighted average indicates the average value weighted according to the number of samples in each achievement level, which better reflects the performance of the model on the overall dataset. Micro-average ROC AUC indicates the value of the area under the ROC curve of the model as a whole, which is an indicator for assessing the overall performance of the model; the closer the value is to 1, the better the performance of the model.

From the classification report, it can be seen that the model's precision and recall rates are excellent for the prediction of each grade interval, especially for the students in the highest and lowest grade intervals, where the model's prediction precision and recall rates reach 100%. This indicates that the model is able to identify students at the extremes of achievement very accurately. For students in the middle achievement intervals, the precision and recall rates also remain high at over 90%, although they decrease slightly. Only the interval with grade 5, which is the group of high-achieving students, has a slightly lower precision rate of 75% but a recall rate of 80%, still indicating that the model's predictions in this interval are acceptable.

In addition, the ROC curve image is shown in Figure 4. The area under the ROC curve (AUC) of the model is 0.971875, which is close to 1. This indicates that the model has an extremely high ability to discriminate between students with different achievement levels. The ROC curve itself also presents a near-ideal condition with a very rapid initial rise, which suggests that the model is able to achieve a high true rate with a low false positive rate. This is a very positive indicator for the prediction model.

Analysis Of Ablation Experiments

In this study, we designed a series of ablation experiments to gain a deeper understanding of the contribution of each base learner in the stacked integration model to the prediction performance. By removing the three base learners—Random Forest, Gradient Boosting Tree, and XGBoost—one by

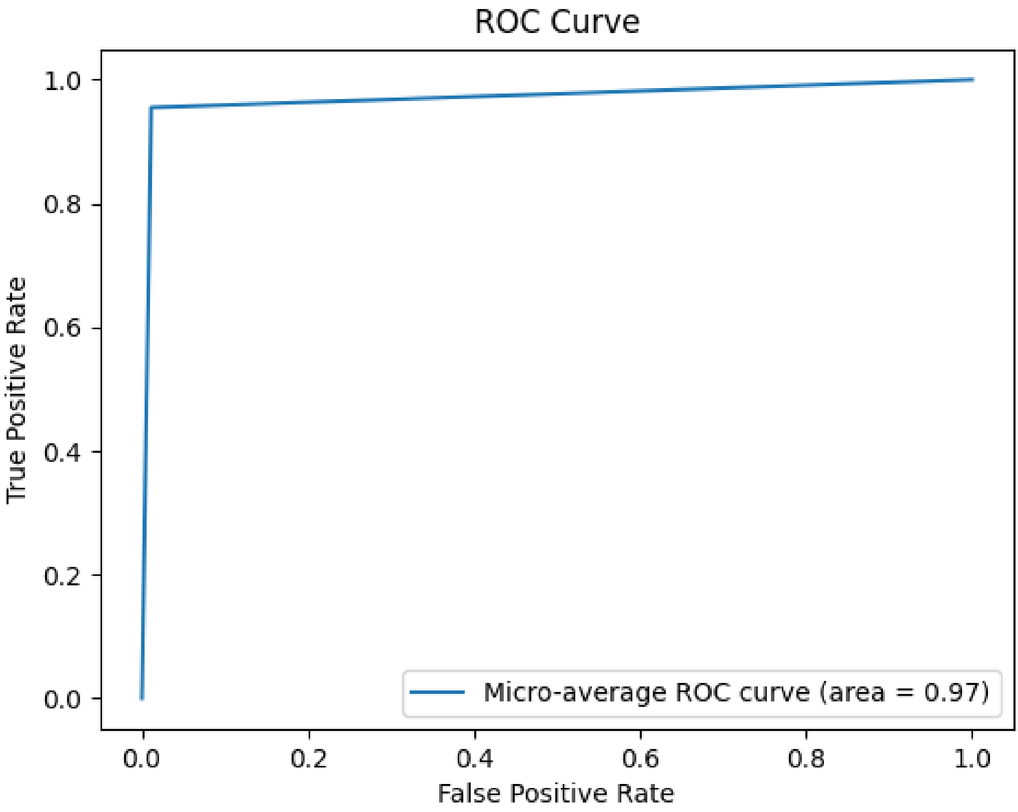
Table 4. Achievement prediction performance of different methods

Experimental group	Removed base learner	Accuracy	Precision (average)	Recall (average)	F1-score (average)
Random Forest + Gradient Boosting Tree + XGBoost	none	0.955	0.93	0.94	0.93
Gradient Boosting Tree + XGBoost	Random Forest	0.935	0.90	0.91	0.90
Random Forest +XGBoost	Gradient Boosting Tree	0.94	0.90	0.91	0.91
Random Forest + Gradient Boosting Tree	XGBoost	0.94	0.90	0.92	0.91

one, we obtained three variants of each model and evaluated the performance of these models. Table 4 summarizes the setup and results of the ablation experiments.

By analyzing the results of the ablation experiments, we found that, after removing the Random Forest base learner, the model accuracy decreased from 0.955 to 0.935, showing the most significant performance degradation. When either the gradient boosting tree or XGBoost is removed, the model accuracy drops to 0.94. Furthermore, in terms of the macro-averaged metrics (precision, recall, and F1 score), the model performance decreases after the removal of either base learner but with a limited

Figure 4. ROC curve image



decrease, suggesting that the stacked integration model has a certain level of robustness and is able to mitigate to some extent the negative impact of individual learner removal.

Overall, our stacked integration model combines multiple behavioral characteristics of students. Through a complex strategy of data preprocessing, feature engineering, and balancing categories, the resulting classifier performs well in predicting student performance. The high accuracy, precision, and recall of the model, coupled with the excellent ROC curve performance, collectively demonstrate the potential and reliability of the model in real-world applications.

DISCUSSION

In this study, we used K-means clustering algorithm, Apriori algorithm, and a stacked integrated model constructed from multiple machine learning models to explore the relationship between student behavioral characteristics and academic performance. Through comprehensive preprocessing and refined analysis of student data, the study reveals a strong correlation between student behavioral characteristics and their academic performance, confirming the value and effectiveness of data-driven methods based on their application within the field of education. Compared with traditional methods, such as the improved genetic algorithm combined with THE BP neural network model, this study not only improves the prediction accuracy by integrating multivariate machine learning techniques but also deepens the understanding of the relationship between behaviors and grades, and it broadens the theoretical and applied horizons of academic performance prediction.

Our model's predictive capabilities can significantly influence educational practices by enabling tailored interventions and personalized learning strategies. By identifying students who are at risk of underperforming early, educators can provide targeted support, enhancing resource allocation, student engagement, and retention rates. For example, students predicted to struggle in specific subjects can receive additional tutoring or alternative learning resources tailored to their needs. Schools can also use these predictions to adjust curriculum delivery, ensuring that teaching methods align more closely with varied student learning styles and capabilities. By discussing practical applications of our findings, we highlight the transformative potential of data-driven approaches in education.

Despite the positive results of this study, there are some limitations. In particular, the study relied on a dataset that was primarily derived from a student population in a specific geographic region and cultural context, which may limit the broad applicability of the findings. In addition, the model performance is largely affected by the quality of data preprocessing and feature selection, and it fails to cover all the factors fully which potentially affect students' academic performance.

To address the above limitations, future research could enhance the generalization ability of the model by collecting data from students from a wider range of geographical and cultural backgrounds. Meanwhile, the introduction of more diverse data sources and the adoption of more advanced feature engineering methods can help to enhance the predictive accuracy and explanatory power of the model further. In addition, future research can explore the complex relationship between student behavioral data and other learning-related variables (for example, mental health status or social network influence), as well as the use of more advanced algorithms such as deep learning, to provide educators with more refined and personalized learning intervention strategies to improve students' academic performance and personal development.

Furthermore, our current model provides a static snapshot of student performance based on behavioral data, but it does not dynamically track changes over time. Recognizing this limitation, future research should explore incorporating longitudinal data analysis to observe and predict trends in student performance more effectively. This enhancement would allow for a continuous assessment of academic risk and the timely adaptation of intervention strategies, thus potentially increasing the model's utility in real-world educational settings. Moreover, expanding our model to include time-series analysis could offer insights into the effects of educational interventions over time, providing a more nuanced understanding of student development and academic progression.

CONCLUSION

This study focuses on exploring the association between students' behavioral characteristics and their academic performance, aiming to improve the accuracy and efficiency of predicting students' academic performance through a data-driven approach. By using K-means clustering algorithm, Apriori algorithm, combined with a stacked integrated model constructed by machine learning models such as Random Forest, Gradient Boosting Tree, and XGBoost, we meticulously analyzed and processed a large amount of student behavioral data. The results of the study show that there is a significant correlation between students' behavioral characteristics, such as frequency of library use, number of times borrowed, length of time spent on the Internet, and classroom and self-study attendance, and their academic performance.

Through in-depth analyses, this study reveals a strong link between specific behavioral patterns and high academic achievement, highlighting the importance of adopting a data-based approach in educational management and pedagogical interventions. In particular, our model achieved an overall accuracy of 95.5% on the test set, demonstrating the potential and reliability of the predictive model in practical applications. This not only provides educators with a powerful tool to predict and improve students' academic performance but also offers new perspectives and approaches for future research.

However, there are limitations to the study, such as the homogeneity of the dataset in terms of geographic and cultural contexts which may limit the broad applicability of the findings. Future studies can overcome these limitations by expanding the diversity of the sample and adopting more advanced data analysis techniques. In addition, the findings also hint at the need to explore the complex relationships between student behavioral data and other learning-related variables (e.g., mental health status, social network influence) further, as well as the possibility of employing more advanced algorithms, such as deep learning, to improve the accuracy of predictive models.

In summary, the findings of this study highlight the importance of understanding and utilizing student behavioral characteristics to improve the quality of education and student academic performance. We expect future research to build on this foundation, not only to improve the accuracy and applicability of the academic performance prediction model but also to explore more factors affecting students' academic performance and provide educators with more comprehensive and effective teaching strategies and interventions. In addition, integrating these findings with smart education platforms and applications may provide students with more personalized and multidimensional learning support, thereby promoting their academic and personal development.

AUTHOR NOTE

The data used to support the findings of this study are included within the article.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

FUNDING INFORMATION

This research was supported by the 2024 Shaanxi University of Science and Technology Higher Education Research Project.

PROCESS DATES

This manuscript was initially received for consideration for the journal on 04/15/2024, revisions were received for the manuscript following the double-anonymized peer review on 05/14/2024, the manuscript was formally accepted on 05/14/2024, and the manuscript was finalized for publication on 05/23/2024

CORRESPONDING AUTHOR

Correspondence concerning this article should be addressed to Jinming Yang, Ulster College, Shaanxi University of Science & Technology, Xi'an, Shaanxi 710021, China. Email: 202115030224@sust.edu.cn.

REFERENCES

Abdul Bujang, S. D., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Md. Ghani, N. A. (2021). Multiclass prediction model for student grade prediction using machine learning. *IEEE Access : Practical Innovations, Open Solutions*, *9*, 95608–95612. 10.1109/ACCESS.2021.3093563

Cao, Z. (2022). Research on prediction and evaluation method of students' performance in colleges and universities based on deep learning. *Information Record Material*, *11*, 123–125. 10.16009/j.cnki.cn13-1295/tq.2022.11.050

Du, X. (2023). *Analysis and application research of student behavior characteristics based on data mining* [unpublished master's dissertation]. Southwest University of Science and Technology., https://link.cnki.net/doi/10.27415/d.cnki.gxngc.2023.000169doi:10.27415/d.cnki.gxngc.2023.000169

Feng, W., Tang, J., & Liu, T. X. (2019). Understanding dropouts in MOOCs. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.

Gao, Y., Wang, J., Li, Z., & Peng, Z. (2023). The social media big data analysis for demand forecasting in the context of globalization: Development and case implementation of innovative frameworks. [JOEUC]. *Journal of Organizational and End User Computing*, *35*(3), 1–15. 10.4018/JOEUC.325217

Hidalgo, Á. C., Ger, P. M., & Valentín, L. D. (2021). Using meta-learning to predict student performance in virtual learning environments. *The International Journal of Research on Intelligent Systems for Real Life Complex Problems, 52*(2022), 3352–3365.

Hu, L. & Zhao, G. (2021). Research on influencing factors of machine learning algorithm on student achievement based on data mining. *Journal of Nanchang University of Aeronautics and Astronautics (Natural Science Edition), 3*, 43-48+97.

Huang, J. (2012). Application of bayesian network to predicting students' achievement. *Computer Science* (S3), *39*(ZII), 280-282.

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, *61*, 133–145. 10.1016/j.compedu.2012.08.015

Li, K. (2023). Grade prediction of university student based on machine learning. *Computing Age*, *12*, 220–223. 10.16644/j.cnki.cn33-1094/tp.2023.12.049

Li, T., & Fu, G. (2010). An overall view of the educational data mining domain. *Modern Educational Technology*, *10*, 21–25.

Liu, A. (2020). The construction of college students performance predictive model based on data mining technology. [Natural Science Edition]. *Journal of Changchun Engineering College*, *2*, 98–101.

Ma, Y. (2020). *Study of college student performance prediction based on machine learning* [unpublished doctoral dissertation]. Shandong University., https://link.cnki.net/doi/10.27272/d.cnki.gshdu.2020.000240doi:10.27272/d.cnki.gshdu.2020.000240

Meier, Y., Xu, J., Atan, O., & Schaar, M. V. (2016). *Personalized grade prediction: A data mining approach* [paper presentation]. *IEEE International Conference on Data Mining*. Barcelona, Spain.

Meng, F., Jiang, S., Moses, K., & Wei, J. (2023). Propaganda information of internet celebrity influence: Young adult purchase intention by big data analysis. [JOEUC]. *Journal of Organizational and End User Computing*, *35*(1), 1–18. 10.4018/JOEUC.318128

Minn, S., Yu, Y., Desmarais, M. C., Zhu, F., & Vie, J. (2018). *Deep knowledge tracing and dynamic student classification for knowledge tracing* [Paper presentation]. *IEEE International Conference on Data Mining*, Singapore.

Pan, F. (2020). *Research on the construction of college students' classification prediction and portrait based on students' daily achievements* [unpublished master's dissertation]. Hebei Agricultural University., https://link.cnki.net/doi/10.27109/d.cnki.ghbnu.2020.000271doi:10.27109/d.cnki.ghbnu.2020.000271

Pardos, Z. A., & Heffernan, N. T. (2010). Modeling individualization in a Bayesian networks implementation of knowledge tracing. In *Proceedings of 18th International Conference on User Modeling, Adaptation, and Personalization*. Springer. 10.1007/978-3-642-13470-8_24

Ren, G., & Yang, Y. (2020). A recurrent neural network based early warning method for students' performance in higher education. *Electronic Technology and Software Engineering*, *15*, 211–212.

Ren, Z., Rangwala, H., & Johri, A. (2016). *Predicting performance on MOOC assessments using multi-regression models* [Paper presentation]. *Ninth International Conference on Educational Data Mining*, Raleigh, NC, USA.

Su, J. (2019). Predictive analysis of student performance in higher education based on data mining. *Information and Communication, 10*, 72-73+75.

Sun, J. (2023). Design of student performance analysis and prediction model based on big data decision tree. *China New Communication*, *9*, 53–55.

Thai-Nghe, N., Horvath, T., & Schmidt-Thieme, L. (2011). *Factorization models for forecasting student performance* [Paper presentation]. *Fourth International Conference on Educational Data Mining*, Eindhoven, The Netherlands.

Wang, H., Zhang, Y., & Jiang, Y. (2024). Research on prediction and analvsis of college students' postgraduate entrance examination results based on data mining. *Journal of Wuyi College*, *1*, 93–97. 10.14155/j.cnki.35-1293/g4.2024.01.014

Wang, Q., Zong, B., Lin, Y., Li, Z., & Luo, X. (2023). The application of big data and artificial intelligence technology in enterprise information security management and risk assessment. [JOEUC]. *Journal of Organizational and End User Computing*, *35*(1), 1–15. 10.4018/JOEUC.326934

Xu, J. (2023). *Analysis and early warning of college students' academic performance based on daily behavior* [unpublished master's dissertation]. North China University of Water Resources and Hydropower., https://link.cnki.net/doi/10.27144/d.cnki.ghbsc.2023.000253doi:10.27144/d.cnki.ghbsc.2023.000253

Yao, M., Li, J., & Wang, N. (2021). Prediction of college students performance based on BP neural network. [Information Science Edition]. *Journal of Jilin University*, *4*, 451–455. 10.19292/j.cnki.jdxxp.2021.04.006

Zeineddine, H., Smith, J. B., & Kumar, A. (2021). Enhancing prediction of student success: Automated machine learning for student data analysis. *Computers in Human Behavior Reports*, *4*, 100130. 10.1016/j.chbr.2021.100130

Zhai, M. (2022). Research on the cause mechanism and intervention strategies of academic crisis based on the university student data [unpublished doctoral dissertation]. Dalian University of Technology, Ganjinzi, Dalian, Liaoning, China. https://link.cnki.net/doi/10.26991/d.cnki.gdllu.2022.003639doi:10.26991/d.cnki.gdllu.2022.003639

Zhang, X., He, X., Du, X., Zhang, A., & Dong, Y. (2023). Supply chain practices, dynamic capabilities, and performance: The moderating role of big data analytics. [JOEUC]. *Journal of Organizational and End User Computing*, *35*(3), 1–26. 10.4018/JOEUC.325214

Zhen, L. (2022). Research on hotspots and frontiers of online education in China in recent five years. *Software Guide*, *9*, 230–235.

Cheng Zhang,Lecturer, Master,Graduated from Shaanxi University of Science & Technology University in 2018. Worked in Shaanxi University of Science & Technology. His research interests include Education management & Computer science.

Jinming Yang, Study in Shaanxi University of Science & Technology University in 2018. His research interests include Machine learning.

Mingxuan Li, Study in Shaanxi University of Science & Technology University in 2018. Her research interests include Education management.

MengDeng,Lecturer, Master,Graduated from Shaanxi University of Science & Technology University in 2018. Worked in Shaanxi University of Science & Technology. Her research interests include Education management.