Complex & Intelligent Systems (2022) 8:323–336
https://doi.org/10.1007/s40747-021-00383-0

ORIGINAL ARTICLE

An interpretable prediction method for university student academic crisis warning

Zhai Mingyu[1] · Wang Sutong[1] · Wang Yanzhang[1] · Wang Dujuan[2]

Received: 31 December 2020 / Accepted: 15 April 2021 / Published online: 29 April 2021

Abstract
Data-driven techniques improve the quality of talent training comprehensively for university by discovering potential academic problems and proposing solutions. We propose an interpretable prediction method for university student academic crisis warning, which consists of K-prototype-based student portrait construction and Catboost–SHAP-based academic achievement prediction. The academic crisis warning experiment is carried out on desensitization multi-source student data of a university. The experimental results show that the proposed method has significant advantages over common machine learning algorithms. In terms of achievement prediction, mean square error (MSE) reaches 24.976, mean absolute error (MAE) reaches 3.551, coefficient of determination (R^2) reaches 80.3%. The student portrait and Catboost–SHAP method are used for visual analysis of the academic achievement factors, which provide intuitive decision support and guidance assistance for education administrators.

Keywords Academic crisis warning · Interpretable machine learning · Student portrait · Catboost–SHAP

Introduction

With the development of informatization in universities, a large amount of data related to student academic performance has been collected, which plays an important role in promoting the education innovation and development. The accumulated big data also provides a good foundation for the application of data-driven techniques in academic warning. More and more scholars pay attention to the enormous social value in educational big data and make research in terms of academic warning. Peterson and Colangelo [1] gave the opinion that boys in colleges were more likely to be in an academic crisis than girls. Reis and McCoach [2] gave a new definition of academic crisis: those who did not meet the standards or the capable ones. It is necessary for students to get required credits within the specified academic years if they want to graduate successfully.

If the credits required for graduation appear to be dropped, the exam should be made up or retaken as soon as possible. The factors of student academic scores deserve the attention of advisors. Advisors are able to adopt various guiding measures to prevent the delay graduation of students in academic crisis if they receive the warning in advance. The credits of students are usually related to study behavior, living behavior, basic information, internet behavior and so on. The data-driven techniques enable university administrators to take fully use of students' data in terms of living habits, family background, etc. Thus, the university administrators and instructors can take timely targeted measures to help students who are at risk of failure to graduate on time or have poor expected performance in next semester. Academic warning based on data-driven techniques is beneficial for discovering the physical or mental health problems of students timely, promoting the all-round development of them, reducing the risk of students delaying graduation or dropping out, better achieving teaching in accordance with their aptitude, and deepening the teaching reform constantly.

Most of the existing methods have low accuracy and interpretability in university student academic crisis warning. They lack the use of living behavior data, internet behavior data for more accurate reflection of students' status. Machine learning methods they used belong to black-box

✉ Wang Dujuan
djwang@scu.edu.cn

[1] School of Economics and Management, Dalian University of Technology, Dalian 116024, China

[2] Business School, Sichuan University, Chengdu 610064, China

methods, which only give the prediction results but cannot provide the inference process. Interpretable machine learning has gradually become a hot topic in academic research in recent years [3]. With the continuous improvement of machine learning method performance, applications in various fields are expanding [4]. However, it is difficult to introduce black-box machine learning methods to some decisions due to the lack of interpretability. It is hard to gain the trust of decision makers without clear reasoning procedure. We need not only accurate but also interpretable methods for academic warning in advance. Student portraits and SHAP-based prediction method are two effective ways to describe the students' conditions and predict the expected academic performance. It is realistic to explore the relationship among study behavior, living behavior, basic information, internet behavior of students. The main contribution of this work is listed as follows:

1. An interpretable prediction method considering categorical features for university student academic crisis warning is proposed, which consists of K-prototype-based student portrait construction and Catboost–SHAP-based academic achievement prediction.
2. A variety of strategies including multi-source data fusion, data filtering, missing value processing, coding transformation are used.
3. Interpretable academic warning visualization consisting of the student portrait and Shapley value plot is realized to give interpretable analysis and provide data-driven decision-making support for university administrators.

The rest parts are stated below. We delineate the related work in terms of academic crisis warning in Section "Related work". Section "An interpretable prediction method considering categorical features" introduces the details of the proposed interpretable prediction method for university student academic crisis warning. We conduct the comparison experiments and give the visualization analysis in Section "Experimental result". Section "Conclusion" concludes our work and give the future direction.

Related work

Traditionally, many scholars carried out the qualitative research on academic crisis warning in higher education in the form of questionnaires, interviews, and surveys. Benjamin and Heidrun [5] explored the relationship between parents' learning ability and children's academic performance. They predicted children's academic performance through parental learning behavior, and found that reducing parental behaviors that were not related to learning could help children improve their academic performance. Barry and Anastasia [6] compared the predictions of students' self-discipline and self-regulation (SR) measures on academic performance, and used multi-source SR questionnaires to identify students' dysfunctions in the process of learning motivation. Fonteyne et al. [7] used questionnaires to explore the factors that affected academic performance, and concluded that in higher education, a suitable learning plan was one of the important factors that promoted the improvement of academic performance. The learning plan was able to better predict academic performance. However, the above methods were easily affected by subjective factors and led to poor generalization performance in different environment.

Recently, more and more scholars tried using data-driven machine learning methods to predict student academic performance. Huang and Fang [8] collected 2907 data from 323 undergraduates in four semesters and used multiple linear regression, multilayer perceptual network, radial basis function network and support vector machine to predict students' scores in the final comprehensive exam. The experimental results showed that support vector machines achieve the highest prediction accuracy. Antonenko and Velmurugan [9] used hierarchical clustering method Wards clustering and non-hierarchical clustering method *k*-means clustering to analyze the behavior patterns of online learners. Dharmarajan and Velmurugan [10] used CHAID classification algorithm to mine information from students' past performance and predict the future performance of students based on the score records of 2228 students. Migueis et al. [11] obtained the dataset of 2459 students from the School of Engineering and conducted comparison results with random forest, decision tree, support vector machine and Naive Bayes. They concluded that random forest is superior to other classification techniques. Yukselturk et al. [12] used machine learning algorithms such as decision tree, K-nearest neighbor, neural networks, and Native Bayes to analyze the causes of dropout. Hachey et al. [13] used a quadratic logistic regression algorithm to analyze the relationship between the students' course notes and academic performance. They concluded that the students' academic performance can be predicted based on the students' course notes. Asif et al. [14] used various data mining methods to predict students' academic achievement and studied typical progressions. Jugo J et al. [15] combined the K-means algorithm with educational data mining to propose an intelligent education and teaching system, which incorporated the design ideas of online games, and improved the final grade of students by allowing students to complete specific tasks. Elbadrawy et al. [16] generated student portraits based on student data, and then used regression analysis and matrix decomposition to predict student performance to help students avoid the risk of failing subjects. Xu et al. [17] predicted undergraduates' academic performance through the Internet behavior by machine learning. The comparison results revealed the association between Internet usage and academic performance.

A large number of experiments on academic crisis warning have been conducted from the qualitative and quantitative perspectives. Data-driven machine learning methods have achieved satisfactory generalization performance [18]. However, there are still many obstacles in the popularization of universities. These methods are black-box methods and cannot provide information about how they achieve predictions. As the ultimate AI user, administrators in universities can only obtain the prediction results, but not the reasons for making specific predictions, which has aroused suspicion and distrust. Only when users can understand why they want to make a specific decision, they will trust them and generate a willingness to use a specific method [19]. Interpretable machine learning presents the internal operating mechanism to users, so that education administrators can not only get more accurate prediction results, but also understand the reasons behind the prediction. At the same time, the possible errors in methods are obvious for users and can be identified and corrected immediately based on the feedback of the education administrators. Frederico et al. [20] attempted to find the factors that affected academic performance through feature importance. They transformed the academic performance prediction into a binary classification problem of whether students successfully completed their studies. They found that the most critical factors affecting performance prediction were the number of courses participated in the school year, the gender of the students and the number of missed subjects using random forest methods. To sum up, there still exists room for improvement in terms of method generalization and interpretability.

An interpretable prediction method considering categorical features

In this paper, we propose an interpretable prediction method considering categorical features for university student academic crisis warning, mainly consisting of K-prototype-based student portrait construction and Catboost–SHAP-based academic achievement prediction. The overall framework of the method is shown in Fig. 1.

For university student big data, it is necessary to perform data preprocessing steps including multi-source data fusion, data filtering, missing value processing, coding transformation, etc. The university big data are mainly made up of two types of features, numerical features including breakfast times in university cafeteria per month, the internet usage time each day etc. and categorical features including gender, birthplace of student, major etc. The two types of features are supposed to be dealt with differently in modeling.

Through early communication with university administrators, we need to first construct the current portrait of the students and then give the prediction academic performance based on the current information. Therefore, we propose K-prototype-based student portrait construction and Catboost–SHAP-based academic achievement prediction. The K-prototype-based student portrait comprehensively describe students from the perspectives of basic information, study behavior, living behavior, and internet behavior. The Catboost–SHAP-based academic achievement prediction gives not only the accurate achievement prediction, but the interpretable feature contribution to the predictions. The interpretable academic warning visualization are presented based on the model output. Thus, an interpretable prediction model for university student academic crisis warning is constructed.

In this paper, we convert academic crisis warning problem into current portrait construction problem and academic performance prediction problem. Based on the dynamic and static data of the students in the T semester, the academic performance of the students in the $T+1$ semester is predicted. Generally, students who are at the bottom of the university or show a significant decline in their grades need academic crisis warning. The judgment threshold is set according to the university conditions.

K-prototype-based student portrait construction

The student portrait represents the common features of the student group, which reflects the specific characters and provides support for student character analysis. The student portrait is usually constructed based on clustering methods.

Clustering is an unsupervised machine learning method that explores the correlation between clusters and evaluates the similarity of data within the cluster. The student portrait is described from the perspectives of basic information etc., similar to the specific student group. Currently popular clustering methods such as K-means, hierarchical clustering, density clustering, etc., can only deal with numerical features. The K-modes algorithm is a clustering algorithm used for categorical feature data in data mining. It is an extension modified according to the core content of K-means, aimed at the measurement of categorical features and the problem of updating the centroid. However, K-modes can only handle categorical feature data. Therefore, there is a need for a clustering method that can process two different types of data at the same time. The K-prototype algorithm inherits the ideas of the K-means algorithm and the K-modes algorithm, and adds a calculation formula describing the dissimilarity between the prototype of the data cluster and the mixed feature data. Considering existence of numerical and categorical features, we cluster the student data based on K-prototype, and build student portraits on the basis of clustering.

In K-prototype algorithm, for numerical features, the Euclidean distance is used. Suppose that the student

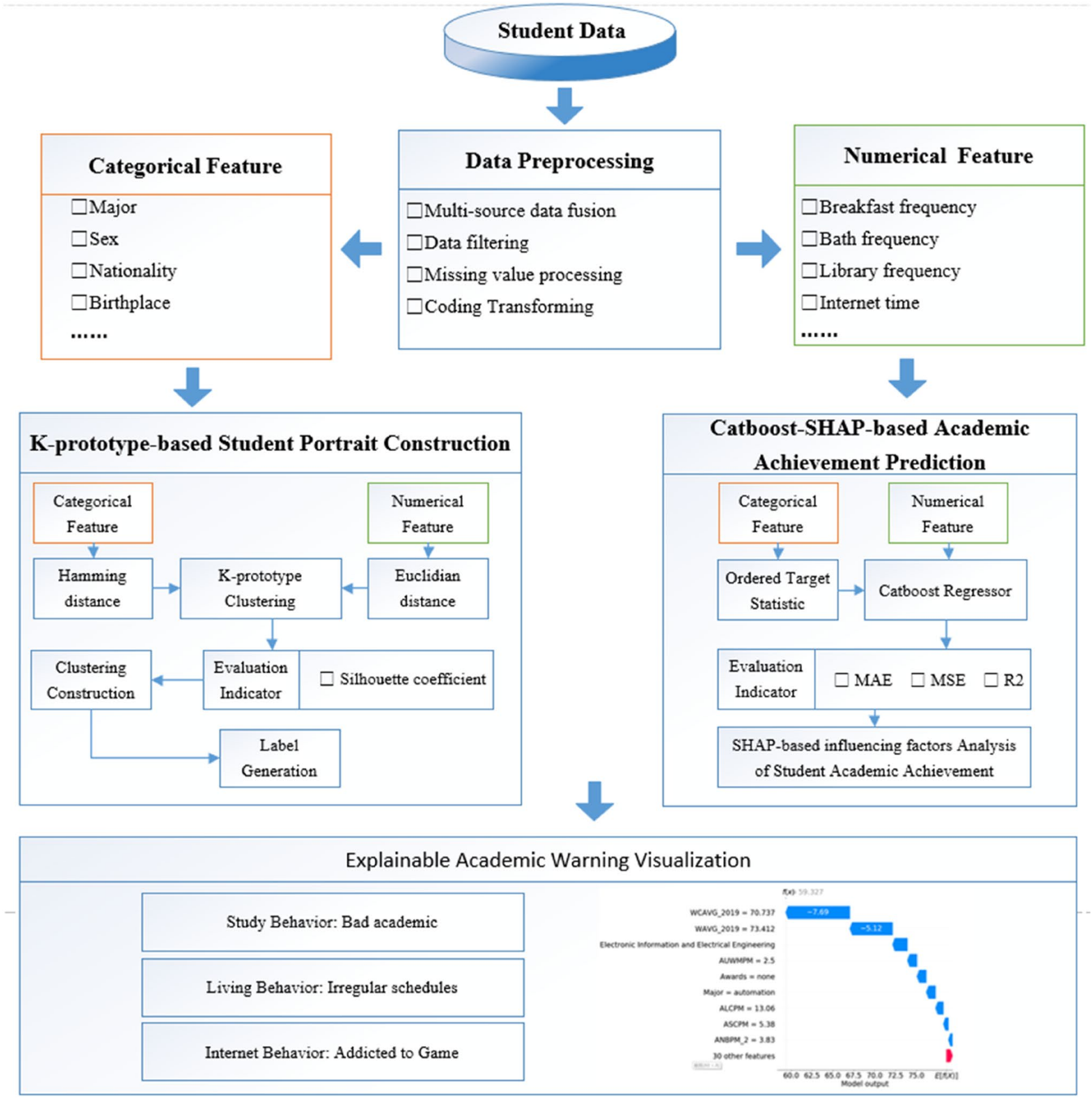

Fig. 1 Framework of the proposed method

dataset with m features and n samples can be expressed with $\boldsymbol{D} = (\boldsymbol{X}_i, y_i) = (\boldsymbol{X}_{\mathbf{num},i} + \boldsymbol{X}_{\mathbf{cat},i}, y_i)$, $i = 1, 2, \ldots, n$. Let $\boldsymbol{X}_{\mathrm{cat},i}$ denotes vector of categorical features and $\boldsymbol{X}_{\mathbf{num},i}$ denotes vector of numerical features, where $\boldsymbol{X}_i \in \boldsymbol{X}$ and $\boldsymbol{X}_i = x_{ij}, j = 1, 2, \ldots, m$. Given two samples $\boldsymbol{X}_a = (\boldsymbol{X}_{\mathbf{num},a} + \boldsymbol{X}_{\mathbf{cat},a})$ and $\boldsymbol{X}_b = (\boldsymbol{X}_{\mathbf{num},b} + \boldsymbol{X}_{\mathbf{cat},b})$. $\boldsymbol{X}_{\mathbf{num},a} = (x_{\mathrm{num},a1}, x_{\mathrm{num},a2}, \ldots, x_{\mathrm{num},am})$ and $\boldsymbol{X}_{\boldsymbol{num},\boldsymbol{b}} = (x_{num,b1}, x_{num,b2}, \ldots, x_{num,bm})$. Student data is first normalized and mapped into the interval [0,1] to reduce the effect of dimensionality. Then Euclidean distance is derived from the distance formula between two points in the Euclidean space and expressed as

$$\mathrm{Euclidean}(\boldsymbol{X}_{\mathbf{num},a}, \boldsymbol{X}_{\mathbf{num},b}) = \sqrt{\sum_{l=1}^{m_{\mathrm{num}}} (x_{\mathrm{num},al} - x_{\mathrm{num},bl})^2}. \tag{1}$$

For categorical features, Hamming distance is calculated. The categorical features part of two samples $\boldsymbol{X}_{\mathbf{cat},a} = (x_{\mathrm{cat},a1}, x_{\mathrm{cat},a2}, \ldots, x_{\mathrm{cat},am})$ and

$\boldsymbol{X}_{\mathbf{cat},\boldsymbol{b}} = \left(x_{\mathrm{cat},b1}, x_{\mathrm{cat},b2}, \dots, x_{\mathrm{cat},bm}\right)$. The expression is listed as follows:

$$\text{Hamming}\left(\boldsymbol{X}_{\mathbf{cat},\boldsymbol{a}}, \boldsymbol{X}_{\mathbf{cat},\boldsymbol{b}}\right) = \sum_{l=1}^{m_{\mathrm{cat}}} \delta\left(x_{\mathrm{cat},al} - x_{\mathrm{cat},bl}\right), \tag{2}$$

where m_{num} and m_{cat} are number of numerical features and categorical features, respectively. If $p = q$, $\delta(p, q) = 0$. If $p \neq q$, $\delta(p, q) = 1$.

The sample dissimilarity of mixed feature types can be calculated through combining different features into a single dissimilarity matrix. Let K be the number of clusters and $Q_c = \left\{q_{c1}, q_{c2}, \dots, q_{cK}\right\}$, which represents the cluster center selected by cluster c, so the distance between the data and the cluster center can be expressed as follows:

$$\text{Distance}\left(\boldsymbol{X}_i, Q_j\right) = \text{Euclidean}\left(\boldsymbol{X}_{\mathbf{num},i}, Q_j\right) + \gamma_c \text{Hamming}\left(\boldsymbol{X}_{\mathbf{cat},i}, Q_j\right). \tag{3}$$

Then, the loss function of K-prototype can be defined as

$$\text{Loss} = \sum_{c=1}^{\mathrm{K}} \left(L_c^{\mathrm{num}} + L_c^{\mathrm{cat}}\right) = L^{\mathrm{num}} + L^{\mathrm{cat}}, \tag{4}$$

L^{num} represents the total loss of all numerical features in the sample of cluster c, L^{cat} represents the total loss of all category features, and γ_c is the weight of categorical features in category c, where γ_c affects the accuracy of clustering. When $\gamma_c = 0$, only numerical features are considered, which is equivalent to the k-means method. The weight of categorical features is greater when γ_c becomes larger, and the clustering result is dominated by categorical features. The proper settings of γ_c results in better cluster performance. It is affected by the mean square error of the numerical variable and is supposed to set 0.5–0.7 when the mean square error is 1. The numerical features are standardized, and the variance is 1, so γ_c is set to 0.5. The specific process of K-prototypes algorithm is shown in Algorithm 1.

We cluster the students from the perspective of living behavior, internet behavior etc. and confirm the number of the target clusters through indicator Silhouette coefficient. After clustering, we further analyze various cluster characteristics and generate character label based on statistics summary of each cluster.

Algorithm 1 K-prototype

Input: Number of target clusters K, Weight factor γ, Iteration T

Output: Clusters C

```
Begin
  For $t < T$:
    Randomly select $K$ initial cluster centers $\{c_1, c_2, \dots, c_K\}$ from the dataset $D$
    Calculate the distance between the sample ($\boldsymbol{X_{num}} + \boldsymbol{X_{cat}}$) and each cluster center $Q_c$
  according to formula (3)
    Assign the sample to the cluster closest to the center $Q_c$
    Update numerical features and categorical features through formulas (1) and (2)
    Use formulas (3) and (4) to calculate the loss function
  Output the clusters $C$
End
```

Catboost–SHAP-based academic achievement prediction

The Catboost–SHAP-based academic achievement prediction is introduced in detail. As a representative of the ensemble learning method, the boosting algorithm has the advantages in prediction accuracy and generalization performance. It continuously adjusts the weight of the sample according to the error rate in continuous iteration, and gradually reduces the deviation of the method. Decision trees are used as base classifiers. The common boosting algorithms such as Adaboost, GBDT do not support the categorical features. The data requires to be transformed with encoding methods such as one-hot encoding before being input to the model, but it performs poorly for the categorical features with high dimensions, which will seriously affect the efficiency and performance effect.

Catboost is an improved version of the boosting algorithm which considers the categorical features. First, the dataset is shuffled, and different permutations are adopted at different gradient boosting stages. By introducing multiple rounds of random permutation mechanism, it effectively improves the efficiency and reduces over-fitting. For a certain value of the categorical feature, it adopts the ordered target statistical (Ordered TS) to deal with the categorical features, which means the categorical feature ranked before the sample is replaced with the expectation of the original feature value. In addition, the priority and its weight are added. In this way, the categorical features are converted into numerical features, which effectively reduces the noise of low-frequency categorical features and enhances the robustness of the algorithm. Suppose the random order of the samples $\rho = (\rho_1, \rho_2, \ldots, \rho_n)$, the sample $x^j_{\rho_U}$ located at j th feature of the sequence ρ_U can be expressed as follows:

$$x^j_{\rho_U} = \frac{\sum_{k=1}^{U-1} I\left(x^j_{\rho_k} = x^j_{\rho_U}\right) \times y_k + a \times U}{\sum_{k=1}^{U-1} I\left(x^j_{\rho_k} = x^j_{\rho_U}\right) + a}, \tag{5}$$

where U is the prior term, and a is the weight coefficient of the prior term greater than 0. On the basis of constructing categorical features, Catboost combines all categorical features, and uses the combined features with higher internal connections as new features to participate in modeling.

Traditional feature importance evaluation methods can only reflect which feature is more important, but cannot show the feature impact on the prediction result. Inspired by the Shapley value of cooperative game theory, the SHAP method [21] constructs an additive interpretation model based on the Shapley value. The Shapley value measures the marginal contribution of each feature to the entire cooperation. When a new feature is added to the model, the marginal contribution of the feature can be calculated with different feature permutations through SHAP.

For student dataset $\boldsymbol{D} = (\boldsymbol{X_i}, y_i)$, the Shapley value of y_i can be expressed as follows:

$$\mathrm{SHAP}(y_i) = E(f(x_{ij})) + \sum_{j=1}^{m} f(x_{ij}), \tag{6}$$

where $f(x_{ij})$ denotes Shapley value of x_{ij} and m corresponds to the number of features. $E(y_i)$ expresses the expected value of all $f(x_{ij})$. When $f(x_{ij}) > 0$, the jth feature of the ith sample has a positive effect on the prediction result y_i, and vice versa, it truly reflects the positive and negative effects of the feature on the prediction result. After deriving the Catboost model, we compute the Shapley values for each feature of dataset. In the training process, the process of constructing the Catboost–SHAP model of a single feature value is shown in Algorithm 2.

First, we input the training data $\boldsymbol{X}$, interested sample x_i, feature j and iteration T. For each iteration, random select a sample z and generate the random permutation of feature. Create two new instances through combining interested x_i and sample z_i. The first interested instance $x_{+\mathrm{j}}$ include x_{j} while x_{j} in $x_{-\mathrm{j}}$ is replaced by permutation z. The feature marginal contribution $f(x^t_i)$ can be calculated through weighted average and output $f(x_i)$. The above steps are repeated for each feature to get the Shapley values for all the features.

Algorithm 2 Shapley additive explanation

Input: dataset $\boldsymbol{X}$, concerned ith sample x_i, iteration number T, Catboost model h, feature j

Output: Shapley value $f(x_i)$

```
Begin
For t=1 to T:
    Random select sample $z_i$ from $\boldsymbol{X}$
    Random select a permutation from feature values
    $x_\tau = (x_1, \dots, x_\mathrm{j}, \dots, x_\mathrm{m})$, $z_\tau = (z_1, \dots, z_\mathrm{j}, \dots, z_\mathrm{m})$
    Combine $x_\tau$ and $z_\tau$ with or without feature $j$
    Sample permutation with feature $j$ $x_{+\mathrm{j}} = (x_1, \dots, x_\mathrm{j}, \dots, z_\mathrm{m})$
    Sample permutation without feature $j$ $x_{-\mathrm{j}} = (x_1, \dots, z_\mathrm{j}, \dots, z_\mathrm{m})$
    $f(x_i^t) = f(x_{+\mathrm{j}}) - f(x_{-\mathrm{j}})$
    $f(x_i) = \frac{1}{T}\sum_{t=1}^{T} f(x_i^t)$
End
```

Experimental result

Data preprocessing

We collect student desensitization data from a university in Dalian, China to conduct experiments. The dataset contains static data such as basic information and dynamic data such as Internet records of students from 2018 to 2020. The details of the dataset can be found Tables 4 and 5.

Data preprocessing accounts for about 80% of the entire workload in data mining, and the quality of data directly affect the performance of model [22, 23]. Therefore, the data needs to be preprocessed before modeling and analysis. Our original dataset comes from multi-source, and there exists problems such as missing data and data redundancy. Data fusion, data filtering, missing value processing, feature code conversion and other data processing steps are required. In data fusion, under the premise of ensuring the integrity of student performance data, the serial number of student is used as the main key to fuse multi-source data.

Feature selection [24] methods have been used in various machine learning methods. We use Random Forest feature selection method to get rid of the useless feature in academic achievement prediction like length of schooling. In this experiment, the original independent features related to academic performance are selected. We screen the student data by academic year and use those of 2018–2019 years as training set and those of 2019–2020 as test set.

According to the domain knowledge related to student management, we compute the monthly average number and consumption of breakfasts, lunches and dinner in the canteen, sports consumption etc. of student consumption record.

For the missing values are less than 10% of the whole dataset, we choose to remain the sample with missing value. In view of the categorical features missing feature values like ethnicity, birthplace, dormitory, loan amount, awards, family economic situation, etc., we fill in uniformly as “none”. In terms of numerical features with missing values like monthly average internet time (h), monthly average internet time at night (h), etc., we fill in with value 0. The weighted average grade (WAVG) is calculated according to the students’ scores and corresponding credits for each academic year according to the following formula:

$$\mathrm{WAVG} = \frac{\sum_{i=1}^{n} \mathrm{grade}_i \times \mathrm{credit}_i}{\sum_{i=1}^{n} \mathrm{credit}_i}. \tag{7}$$

In the process of K-prototype-based student portrait construction, after missing data filtering, we use maximum and minimum normalization to deal with numerical features. We use the

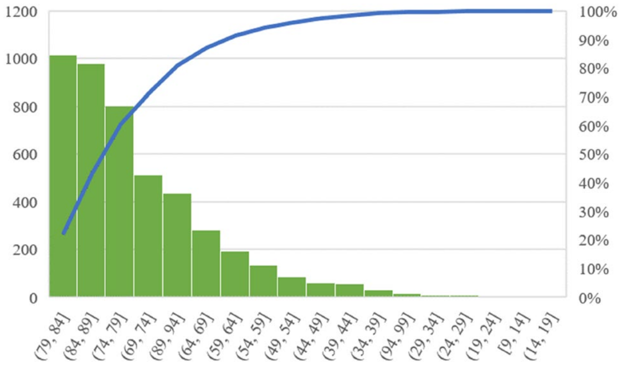

Fig. 2 Cumulative distribution of student academic performance for 2017 grade student

following formula to normalize the numerical features of each sample to reduce the impact of different feature distances:

$$X_{ij}^{*} = \frac{X_{ij} - X_{\min}}{X_{\max} - X_{\min}}, \tag{8}$$

where X_{ij} and X_{ij}^{*} denote the value before and after normalization. X_{mean} and X_{std} correspond to the mean value and standard deviation of the feature.

Data description

After data preprocessing, a total of 13,613 student data are obtained. We select 4,624 student samples of 2017 grade because the compulsory courses of the second year and the third year are more comprehensive. The data can be described from four perspectives including the basic information, study behavior, internet behavior, and living behavior.

Basic information includes the description of student such as gender, ethnicity, date of birth, family structure, admission type, birthplace and family economic status. The study behavior mainly includes the weighted average grades and the failed grades of the previous academic year, the number of visits to the library, the number of borrowed books, the information of the student's department, major, class, the number of awards, and the amount of scholarship loans. Internet behavior mainly include monthly average internet time (h), monthly average internet time at night (h), network traffic usage, game online time, the number of commonly used APPs, etc. Living behavior refers to a way of activity and configuration of students, which mainly contains the monthly average number and consumption of breakfasts, lunches and dinner in the canteen, sports consumption, frequency of water usage, frequency of bathing, frequency of washing machine use, time for returning to the dormitory every night etc. The 2017 grade student samples are listed in Tables 4 and 5 according to the numerical features and categorical features.

The data in Tables 4 and 5 reflect the overall performance of the 2017 grade students in terms of study and life. When analyzing performance of a single student, it can be combined with the overall situation of the school for research and exploration.

The histogram in Fig. 2 reflects the overall distribution of student scores in the 2018–2019 academic year of the university. From Fig. 2, it can be seen that the proportion of students with weighted average grade in the 79–84 intervals ranks first. The line chart reflects the cumulative changes in each performance interval. The weighted average grade in the 60–94 intervals accounts for 95% of the overall ratio. We set 60 as the threshold of crisis warning as the students with the weighted average grade below 60 rank around the last 5% of all the students and deserve the additional attention of administrators.

Performance metrics

To validate the performance of K-prototype-based student portrait construction, the Silhouette coefficient, Calinski-Harabasz and Davies Bouldin score are used. The Silhouette Coefficient combines the cohesion and separation to evaluate the clustering performance. The formula of Silhouette Coefficient is shown as follows:

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{g_i - v_i}{\max\left\{g_i, v_i\right\}}, \tag{9}$$

where v_i represents the cohesion of cluster, which means the average distance among the ith sample and all other data in the same cluster. g_i represents the separation, which means the distance between the ith sample and the nearest cluster.

Table 1 Comparative results of clustering performance

Models	Cluster	S	CH	DBI
K-means	2	0.428484	7095.454	0.892379
	3	0.398637	6153.234	0.970542
	4	0.408408	6160.945	0.858285
	5	0.389156	5622.735	0.933592
	6	0.331472	5579.942	0.933686
	7	0.33504	5557.057	0.951496
	8	0.316842	5456.002	1.025409
	9	0.277598	5007.052	1.079333
	10	0.269662	4748.582	1.19619
Birch	2	0.360267	5495.627	0.805574
	3	0.323743	5318.713	0.99023
	4	0.382148	5594.193	0.86904
	5	0.331424	5358.336	0.94342
	6	0.319297	5317.621	1.010787
	7	0.334224	5164.199	1.016429
	8	0.325813	5093.434	0.991862
	9	0.335003	5113.016	0.988204
	10	0.328125	5086.486	1.021491
MeanShift	–	0.472562	6257.606	**0.692773**
OPTICS	–	– 0.17052	16.7709	1.548755
K-prototype	2	**0.496154**	**7396.385**	0.732036
	3	0.424015	7149.989	0.88925
	4	0.415818	6278.954	0.912406
	5	0.407517	6164.507	0.843537
	6	0.370032	6079.004	0.921779
	7	0.35086	5882.694	0.958512
	8	0.349542	5773.671	0.931606
	9	0.344894	5583.745	0.996182
	10	0.332636	5454.374	0.993635

Bold values indicate better results than other filtering methods

When $S < 0$ and $g < v$, the clustering performance is not good. When v_i tends to 0, or g is much larger than v, S tends to 1, which means the model achieves a good performance.

Calinski–Harabaz Index is expressed as follows:

$$\mathrm{CH} = \frac{\mathrm{Tr}\left(B_k\right)}{\mathrm{Tr}\left(W_k\right)} \times \frac{N-k}{k-1}, \tag{10}$$

where B_k denotes between-clusters dispersion mean and W_k corresponds to within-cluster dispersion. When the covariance of the data within the cluster is smaller and the covariance of the data between the clusters is larger, the performance of the method will be better, which means that the larger the CH index value is, the better the performance of the model will be.

Davie Bouldin Score is shown as follows:

$$\mathrm{DBI} = \frac{1}{n}\sum_{i=1}^{n} \max\left(\frac{s_i - s_j}{\left\|w_i - w_j\right\|_2}\right), \tag{11}$$

where s_i indicates the degree of dispersion of data points in the ith cluster. The minimum value of DBI is 0, and the smaller the value is, the better the clustering effect is.

For the evaluation of Catboost–SHAP-based academic achievement prediction, we use the common performance indicators of regression methods, such as mean square error (MSE), mean absolute error (MAE) and coefficient of determination (R^2) [25]. Assuming that n is the number of samples, y_i^{pred} is the predicted value of the i th sample, y_i and $\overline{y}$ denote the corresponding true value, respectively. Then the three indicators can be expressed as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i^{\text{pred}}\right)^2 \tag{12}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\left(y_i - y_i^{\text{pred}}\right)\right| \tag{13}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_i^{\text{pred}}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}. \tag{14}$$

Table 2 Parameter settings of Catboost–SHAP

Parameter	Default value	Improved value
Number of iterations	1000	9000
Learning rate	0.03	0.1
Maximum depth	6	10
Maximum One hot size	2	2
Categorical features	None	$\boldsymbol{X}_{\mathbf{cat}}$
Loss function	RMSE	MSE
L2 leaf regularization	0	3
Device	CPU	GPU

Performance comparison

Comparison results of K-prototype-based student portrait construction

We compare the K-prototype clustering method with popular clustering methods including K-means, Birch, MeanShift, OPTICS and use Silhouette Coefficient, Calinski-Harabasz and Davies Bouldin score to analyze the performance under different clusters. We conduct the experiments on the whole dataset and the comparison is shown in Table 1. Birch, MeanShift, OPTICS do not need to set the number of clusters and we mark '–' for distinction.

It can be seen from Table 1 that K-prototype performs significantly better than other clustering methods in terms of Silhouette coefficient and Calinski-Harabasz. K-prototype have the best performance in terms of various indicators when the number of clustering is set 2 for all the dataset. MeanShift performs better in terms of Davies Bouldin score. It reflects K-prototype clustering is more effective when data contains both categorical and numerical features. Through K-prototype, students can be divided into different clusters and labeled with different tag from the view of living behavior, study behavior and Internet behavior. In addition, the single student shares the common characters of the student group.

Comparison results of Catboost–SHAP-based academic achievement prediction

To test the performance of the Catboost–SHAP method in regression prediction, we have the experiments with our

Fig. 3 Relationship of the loss versus iterations of Catboost–SHAP

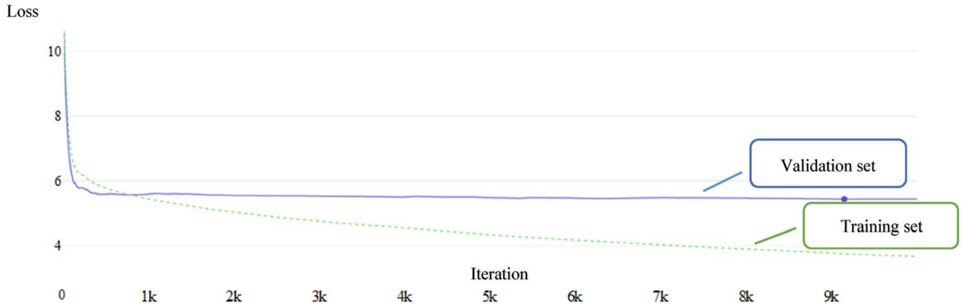

Table 3 Performance comparison of student academic prediction methods

Method	Prediction Time	MSE	MAE	R^2
KNN	0.026 (±0.001)	80.485 (±12.223)	6.464 (±0.181)	0.366 (±0.061)
LR	0.007 (±0.001)	42.734 (±10.354)	4.471 (±0.132)	0.665 (±0.058)
DT	0.132 (±0.005)	43.143 (±9.735)	4.380 (±0.144)	0.661 (±0.056)
SVM	**0.005** (±**0.000**)	90.636 (±16.353)	6.214 (±0.214)	0.288 (±0.096)
MLP	0.237 (±0.001)	133.200 (±10.768)	8.037 (±**0.109**)	– 0.051 (±0.018)
RF	0.006 (±**0.000**)	47.968 (±9.824)	4.774 (±0.184)	0.623 (±0.057)
BAG	0.174 (±0.003)	42.950 (±9.686)	4.381 (±0.139)	0.663 (±0.055)
ADB	0.083 (±0.029)	61.522 (±10.972)	6.024 (±0.381)	0.516 (±0.064)
GBDT	0.010 (±0.005)	41.236 (±10.103)	4.258 (±0.131)	0.676 (±0.058)
XGBoost	0.013 (±0.001)	40.785 (±10.334)	4.240 (±**0.109**)	0.680 (±0.058)
LightGBM	0.008 (±**0.000**)	41.177 (±10.084)	4.254 (±0.131)	0.677 (±0.057)
Catboost–SHAP	0.657 (±1.096)	30.254 (±6.749)	3.723 (±0.162)	0.763 (±**0.03**)
Improved Catboost–SHAP	0.061 (±0.006)	**24.976** (±**5.941**)	**3.551** (±0.162)	**0.803** (±0.034)

Bold values indicate better results than other filtering methods

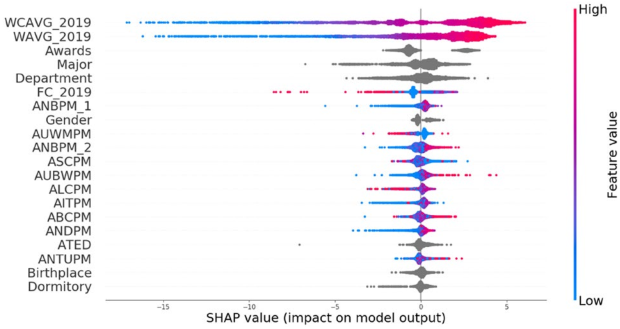

Fig. 4 Feature importance ranking plot with improved Catboost–SHAP

proposed method and other popular machine learning methods such as Linear regression (LR), support vector machine (SVM), decision tree (DT) and commonly used ensemble learning methods adaptive enhancement (AdaBoost), random forest (RF), gradient boosting decision tree (GBDT), XGBoost, LightGBM for comparison. To validate the generalization of our proposed method, tenfold cross validation is used, and each comparison experiment was carried out ten times independently to ensure the validity of the experiment.

We train the comparative method on student data of 2018–2019 academic year and perform prediction on the weighted average grade (WAVG) of 2019–2020 academic year. For the parameter setting of Catboost–SHAP, we adopt the default settings to compare with other methods and separate the validation set from the training set to further improve the performance of Catboost–SHAP. To check the model convergence effect, we plot the relationship of the loss versus iterations of Catboost–SHAP in Fig. 3.

In Fig. 3, the green dotted line represents the loss decreasing with iterations of training set and the blue solid line denotes the loss decreasing with iterations of validation set. The best performance of validation set is around 9000 iterations, represented by the blue dot in the figure. Therefore, we adopt 9000 iterations and tune the other parameters through grid search method. The default value of original settings of Catboost–SHAP and the best parameters settings of improved version of Catboost–SHAP are shown in Table 2.

To make a fair comparison with other methods, we use default parameters for all methods including Catboost–SHAP. To validate the effectiveness of the improved Catboost–SHAP, we add it to the comparison results and the comparative experimental results are shown in Table 3.

We compare the mean and variance of performance indicators of various methods over tenfolds. The results in Table 3 show that the Catboost–SHAP proposed is superior to other methods in terms of MSE, MAE and R^2. Catboost–SHAP achieves the smallest value in MSE, MAE and realize the largest value in R^2, which shows the excellent fitting ability.

To further improve the performance of Catboost–SHAP, we optimize the parameter settings, tune the parameters as Table 2 and achieves better performance compared with original one, which achieves 17.45% improvement in MSE, 4.63% in MAE and 5.26% in R^2. In addition, it costs shorter prediction time with the help of GPU device. It has the smallest variance in MSE in the tenfold cross validation.

Compared with other popular methods, the prediction time of Catboost–SHAP is slightly longer, but it is at the millisecond level, which has no significant difference.

Interpretable analysis

To ensure the generalization ability and stability of the prediction, it is significant to find the core factors that affect student academic performance based on the student portrait and the prediction results. The analysis based on portrait and SHAP go deep into the model to give a reasonable explanation for the prediction results. It tells the teacher which

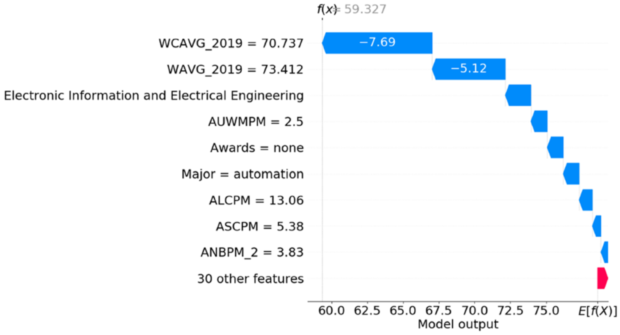

Fig. 5 Shapley value plot of the student

aspect of the students need to pay more attention to, what are the reasons for the poor grades or missed subjects, so as to provide targeted guidance to the students.

We calculate the Shapley value of all student data with Catboost–SHAP-based academic achievement prediction and draw a feature importance ranking plot in Fig. 4.

Figure 4 plots the SHAP value of each feature for all samples. Each row represents a feature, and the abscissa corresponds to the SHAP value. Each point in the plot represents a sample, where red represents positive contribution and blue represents negative contribution. The absolute mean values of Shapley are calculated for each feature and are sorted from top to bottom to represent the rank of feature importance. According to the order, the weighted average grades in the previous academic year, the weighted compulsory average grades in the previous academic year, awards, major, department, failed credits in the previous academic year, dormitory make sense to the academic performance prediction. The red part of figure indicates that WAVG_2019, WCAVG_2019, etc. are proportional to the final score. The increase in the value of these features can improve the predicted scores, while the blue part like FC_2019, AUBWPM, ANBPM_1 are inversely proportional to the final score. From the features, it can be seen that the scores in the previous academic year account for a large proportion of the forecast. In addition, awards, major, the dormitory atmosphere, breakfast time and good reading habits are very important for getting good grades. Through the plot, we can better understand the internal operating mechanism of the prediction model, enhance the trust of education administrators.

Case study with interpretable academic warning visualization

We have performed the K-prototype-based student portrait construction on the student dataset from the perspective of study behavior, living behavior and internet behavior. We define the clusters referenced to the statistics summary of all the students. From the study behavior perspective, the students are divided into 4 groups, including bad academic, medium academic, good academic and excellent academic. In terms of living behavior, 3 clusters are generated, including extremely irregular schedules, irregular schedules, regular schedules. The internet behavior can be transferred to addicted to game, normal internet usage, seldom internet access. The student sample belongs to bad academic in the study behavior, irregular schedules in living behavior and addicted to game in the internet behavior.

We present the analysis results of the Catboost–SHAP model on academic performance. With the help of visualization, the internal operation mechanism of the Catboost–SHAP model can be explored. A student who needs academic crisis warning is listed in Fig. 5 as example for empirical research.

The red and blue in Fig. 5 show the positive and negative contributions of each feature to the final prediction score, pushing the model's prediction results from the basic value to the final value. The basic value is the mean value of the model prediction on the test set. The WCAVG_2019 is 70.737, the WAVG_2019 is 73.412. The mean grades of department of electronic information and electrical engineering is generally lower than other department, which means the harder level of courses. His average usage of washing machine per month (AUWMPM) is 2.5, which is higher than the average level, which indicates more time in dormitory. Through the visualization plot, we can know the internal mechanism of the model's prediction, which is easier for education administrators to understand.

Conclusion

Academic crisis warning of university students enable administrators to pay attention to students' academic problems as early as possible. The student portrait and accurate academic performance prediction give interpretable analysis and provide data-driven decision-making support for university administrators. In our study, the 2018–2020 desensitized student data of a university in Dalian, China is used for prediction experiments. After preprocessing of multi-source data, it is input into our proposed framework with K-prototype-based student portrait construction and Catboost–SHAP-based academic achievement prediction for university student academic crisis warning. It gives high-performance machine learning methods with visual interpretability analysis, and in-depth exploration of students' daily life, study habits on the basis of achieving academic early warning. The student portrait and relationship between factors and academic performance provide guidance assistance and decision support for university administrators and instructors. We train our interpretable prediction method based on the actual student data after desensitization in a university, and compare the method with other mainstream machine learning methods. The experimental results show that our method has significant advantages in the performance and performance of the method, which is better than machine learning LR, DT, SVM, RF, BAG, ADB, GBDT, XGBoost, LightGBM in the method. In tenfold cross validation, the MSE of the Catboost–SHAP method is 24.976, the MAE is 3.551, and the R^2 is 80.3% in terms of academic performance prediction.

Student academic crisis warning of students based on our method can detect problematic students with poor expected grades as early as possible, and can also analyze specific factors that are positively and negatively related to their grades. Good course scores in last academic year, regular living habits all reflect a positive correlation with greater weight. Through interpretable academic warning visualization, we can further analyze the reasons behind their poor performance and provide timely guidance and suggestions for university administrators.

In future research work, we will consider incorporating more time-series dimensional data to conduct in-depth mining from a more comprehensive view. At the same time, we will consider integrating more educational data from other sources and realize a more real time, accurate and stable student academic crisis warning, which provide more comprehensive decision-making support for education administrators.

Appendix

See Tables 4 and 5.

Table 4 2017 grade student numerical features

Feature type	Numerical feature	Feature description	Mean	Std	Median	Maximum
Study behavior	WCAVG_2019	Weighted compulsory average grades in the previous academic year	76.73	11.70	79.41	96.00
	FC_2019	Failed credits in the previous academic year	5.67	11.50	0.00	127.50
	WAVG_2019	Weighted average grades in the previous academic year	76.97	10.66	79.32	96.00
	NLEPM	Number of library entries per month	2.47	3.91	1.10	64.20
	BBPM	Borrowed books per month	0.33	0.92	0.00	21.00
Living behavior	ANBPM_1	Average number of breakfasts per month in the cafeteria during breakfast time (5–10 o'clock)	7.42	5.03	6.38	28.00
	ABCPM	Average breakfast consumption per month in the cafeteria during breakfast time (5–10 o'clock)	5.96	1.86	5.71	24.05
	ANLPM	Average number of lunches per month in the cafeteria during lunch time (10–15 o'clock)	9.07	5.14	8.50	32.00
	ALCPM	Average lunch consumption per month in the cafeteria during lunch time (10–15 o'clock)	11.46	2.06	11.38	27.04
	ANDPM	Average number of dinners per month in the cafeteria during dinner time (15–20 o'clock)	7.86	4.81	7.21	33.50
	ABDPM	Average number of dinners per month in the cafeteria during dinner time (15–20 o'clock)	10.93	2.29	10.93	27.14
	AUWMPM	Average usage of washing machine per month	0.42	1.04	0.00	16.92
	ANBPM_2	Average number of baths per month	4.08	3.44	3.42	21.83
	AUBWPM	Average usage of boiling water per month	12.80	13.15	9.75	135.50
	ANSPM	Average number of sports per month in the gym	0.43	0.81	0.08	14.08
	ANHVPM	Average number of hospital visits per month	0.02	0.07	0.00	1.25
	AHCPM	Average hospital consumption per month	3.99	11.26	0.00	175.45
	ASCPM	Average supermarket consumption per month	3.74	4.10	2.63	63.92
	ANBRPM	Average number of school bus rides per month	0.12	0.35	0.00	5.71
Internet behavior	AITPM	Average Internet time per month (h). If there are multiple connected devices to WLAN, the time is accumulated	293.85	225.04	268.47	1475.41
	AITNPM	Average Internet time at night per month (h) (0–6 o'clock). If there are multiple connected devices to WLAN, the time is accumulated	9.84	12.12	5.61	97.75
	ANTUPM	Average network traffic (GB) usage per month. If there are multiple connected devices to WLAN, the traffic is accumulated	36.21	30.63	31.80	253.60
	AOTOEA	Average online time of once entertainment apps (min)	30.94	25.69	28.12	334.68
	NEA	Number of entertainment apps	5.03	3.16	5.00	19.00
	MTEA	Maximum time of entertainment APP (min)	234.65	255.81	157.71	1439.98

Table 5 2017 grade student categorical features

Feature type	Categorical feature	Feature description	Type number	Type sample
Basic information	Gender	Reflects the gender differences	2	Male, Female
	Ethnicity	Reflects ethnic differences	31	Han, Hui
	Family_structure	Reflect single parent family or not and the influence of family	3	Single
	Admission_type	Reflects the differences among students of different types of admission, such as differences between urban and rural areas, etc	9	Rural fresh
	Birthplace	Reflect differences in habitats	33	Liaoning, Heilongjiang
	Family_economic_status	The degree of difficulty reflects the differences in the status of different families	3	Normal, Especially difficult
Study behavior	Department	Reflect the differences of different departments	21	School of economic and management
	Major	Majors reflect the differences of different majors	83	Philosophy, business administration
	Dormitory	The name of the dormitory reflects the difference in dormitory learning style	26	13th dormitory, 14th dormitory
	Awards	Number of awards Scholarships and awards can reflect students' club activities and learning	3	1 time, 2 times
Living behavior	ATED	Average time entrance into the dormitory	16	16 h, 17 h
	Loan_amount	The loan amount reflects the student's family situation	20	14,000 CNY, 15,000 CNY
	Funding	Reflects the student's family situation	5	2000 CNY, 3000 CNY
Internet behavior	HFEA	High-frequency entertainment APP which reflects the leisure and entertainment APP used most frequently	36	King of Glory

Acknowledgements This paper is our original work and has not been published or submitted simultaneously elsewhere. All authors have agreed to the submission and declared that they have no conflict of interest. This paper was supported in part by the National Natural Science Foundation of China (No. 71533001).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Peterson JS, Colangelo N (1996) Gifted achievers and underachievers: a comparison of patterns found in school files. J Couns Dev 74:399–407. https://doi.org/10.1002/j.1556-6676.1996.tb01886.x
2. Reis SM, McCoach DB (2000) The underachievement of gifted students: what do we know and where do we go? Gift Child Q 44:152–170. https://doi.org/10.1177/001698620004400302
3. Preece A (2018) Asking "Why" in AI: explainability of intelligent systems—perspectives and challenges. Intell Syst Accounting, Financ Manag 25:63–72. https://doi.org/10.1002/isaf.1422
4. Aslam M (2019) Neutrosophic analysis of variance: application to university students. Complex Intell Syst 5:403–407. https://doi.org/10.1007/s40747-019-0107-2
5. Matthes B, Stoeger H (2018) Influence of parents' implicit theories about ability on parents' learning-related behaviors, children's implicit theories, and children's academic achievement. Contemp Educ Psychol 54:271–280. https://doi.org/10.1016/j.cedpsych.2018.07.001
6. Zimmerman BJ, Kitsantas A (2014) Comparing students' self-discipline and self-regulation measures and their prediction of academic achievement. Contemp Educ Psychol 39:145–155. https://doi.org/10.1016/j.cedpsych.2014.03.004

7. Fonteyne L, Duyck W, De Fruyt F (2017) Program-specific prediction of academic achievement on the basis of cognitive and non-cognitive factors. Learn Individ Differ 56:34–48. https://doi.org/10.1016/j.lindif.2017.05.003
8. Huang S, Fang N (2013) Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models. Comput Educ 61:133–145. https://doi.org/10.1016/j.compedu.2012.08.015
9. Antonenko PD, Toy S, Niederhauser DS (2012) Using cluster analysis for data mining in educational technology research. Educ Technol Res Dev 60:383–398. https://doi.org/10.1007/s11423-012-9235-8
10. Dharmarajan A, Velmurugan T (2013) Applications of partition based clustering algorithms: a survey. In: 2013 IEEE International Conference on computational intelligence and computing research. IEEE, pp 1–5
11. Miguéis VL, Freitas A, Garcia PJV, Silva A (2018) Early segmentation of students according to their academic performance: A predictive modelling approach. Decis Support Syst 115:36–51. https://doi.org/10.1016/j.dss.2018.09.001
12. Yukselturk E, Ozekes S, Türel YK (2014) Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. Eur J Open, Distance E-Learning 17:118–133. https://doi.org/10.2478/eurodl-2014-0008
13. Hachey AC, Wladis CW, Conway KM (2014) Do prior online course outcomes provide more information than G.P.A. alone in predicting subsequent online course grades and retention? An observational study at an urban community college. Comput Educ 72:59–67. https://doi.org/10.1016/j.compedu.2013.10.012
14. Asif R, Merceron A, Ali SA, Haider NG (2017) Analyzing undergraduate students' performance using educational data mining. Comput Educ 113:177–194. https://doi.org/10.1016/j.compedu.2017.05.007
15. Jugo I, Kovačić B, Slavuj V (2016) Increasing the adaptivity of an intelligent tutoring system with educational data mining: a system overview. Int J Emerg Technol Learn 11:67. https://doi.org/10.3991/ijet.v11i03.5103
16. Elbadrawy A, Polyzou A, Ren Z et al (2016) Predicting student performance using personalized analytics. Computer (Long Beach Calif) 49:61–69. https://doi.org/10.1109/MC.2016.119
17. Xu X, Wang J, Peng H, Wu R (2019) Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Comput Human Behav 98:166–173. https://doi.org/10.1016/j.chb.2019.04.015
18. Lu J, Liu A, Song Y, Zhang G (2020) Data-driven decision support under concept drift in streamed big data. Complex Intell Syst 6:157–163. https://doi.org/10.1007/s40747-019-00124-4
19. Ribeiro MT, Singh S, Guestrin C (2016) "Why should i trust you?" In: Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining. ACM, New York, NY, USA, pp 1135–1144
20. Cruz-Jesus F, Castelli M, Oliveira T et al (2020) Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. Heliyon 6:e04081. https://doi.org/10.1016/j.heliyon.2020.e04081
21. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in neural information processing systems
22. García S, Luengo J, Herrera F (2016) Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowl-Based Syst 98:1–29. https://doi.org/10.1016/j.knosys.2015.12.006
23. Wang S, Wang Y, Wang D et al (2020) An improved random forest-based rule extraction method for breast cancer diagnosis. Appl Soft Comput 86:105941. https://doi.org/10.1016/j.asoc.2019.105941
24. Hoque N, Singh M, Bhattacharyya DK (2018) EFS-MI: an ensemble feature selection method for classification. Complex Intell Syst 4:105–118. https://doi.org/10.1007/s40747-017-0060-x
25. Boodhun N, Jayabalan M (2018) Risk prediction in life insurance industry using supervised learning algorithms. Complex Intell Syst 4:145–154. https://doi.org/10.1007/s40747-018-0072-1