Achievement Performance Prediction Model Based on Deep Neural Network of Student Similarity

Yixuan Rong[1], Tingnian He[2*], Zhuoran Li[3], Guoqi Liu[4]

{2021222260@nwnu.edu.cn[1], 87956426@163.com[2*], 2021222205@nwnu.edu.cn[3], 1932349928@qq.com[4]}

College of Computer Science & Engineering, Northwest Normal University, Lanzhou, China

Abstract. Student performance prediction aims to predict performance through student information and guide students and teachers to optimise the learning process. Traditional methodological studies focus more on students' own influencing factors and ignore the similarity of different students' learning abilities. To address this problem, the performance prediction model based on Deep Neural Network of student similarity (Sim-DNN) takes into account student-to-student associations, and predicts students' academic performance by calculating the similarity of different students' attribute features, selecting students with higher similarity to the target students, and weighting and summing the historical scores of the similar students according to the degree of similarity as the input to the deep neural network . In order to minimise the effect of overfitting, the model incorporates Dropout in the DNN. The experimental results show that the model proposed in this paper has a better prediction performance on both public datasets, Mathematics and Portuguese.

Keywords: Achievement prediction; Deep Neural Networks; Dropout

1 Introduction

With the continuous development of Internet technology, information technology has invariably changed people's way of life. In this context, all kinds of information technology push forward, bringing brand-new opportunities to the field of intelligent education. Smart education is different from traditional education in that it can better recognize individual differences and different needs through the application of modern information technology, the Internet, big data, and artificial intelligence in education. How we can utilize technologies such as data mining, machine learning and deep learning to provide educators with valuable information to more scientifically improve the way of education and teaching and to make decisions on management services, and thus to obtain better teaching effectiveness and educational outputs, is a very worthwhile issue to be studied. Intelligent education is the inevitable choice and development form of education in the information age to meet the needs of social development.

The prediction of students' academic performance has received attention from many experts and scholars as an important part of cognizing students' learning in smart education. By predicting students’ performance, we can identify the risk of students’ academic failure in the learning process as early as possible, so as to intervene and guide in advance[1]. Specifically,

ICMEIM 2023, September 08-10, Wuhan, People's Republic of China
Copyright © 2023 EAI
DOI 10.4108/eai.8-9-2023.2340180

the student performance prediction model is based on the analysis and prediction of students' past performance such as historical grades, study situation, elective courses, etc., and the analysis of possible influencing factors such as family situation, health condition, learning environment, and online behavior. The model is analyzed to predict whether the target student will perform well in the next grade level or GPA ranking for the purpose of academic alert. This allows teachers to react in time and adjust their teaching strategies and management style. This allows students to have a higher degree of learning freedom; students are free to set their plans and goals, which also leads to more factors that affect academic performance[2]. In previous studies, researchers have more often used machine learning algorithms to enhance the performance of predictive models. Francis et al[3] classified the characteristic factors affecting students' performance into four categories: demographic, academic, behavioral, and additional characteristics. It proposed a method combining classification and clustering to predict performance, and the results showed that the best prediction results were obtained when academic, behavioral, and additional characteristics were considered together. Ghorbani et al[4] compared multiple resampling techniques to predict student dropout rates using two datasets to deal with the data imbalance problem while improving the performance of the predictive model. Experimentally, the combination of Random Forest classifier and SVM-SMOTE balancing algorithm performed best on the dataset.

Despite the progress that has been made in the above studies, there is still a lot of room for improvement in predictive modeling of student performance. Existing prediction models are mostly based on classical models of machine learning, especially those related to supervised learning, where model performance needs to be improved and student-to-student connections are ignored. Deep learning has excellent performance in other domains, but it has less application in student performance prediction and needs to be further explored. Therefore, this paper proposes achievement performance prediction model based on deep neural network of student similarity. In order to delve deeper into the prediction model, this paper analyzes and mines two publicly available datasets to explore the factors that influence student performance. Finally, the model of this paper is applied on the public dataset and compared and analyzed with the existing algorithms, and the experimental results show that the accuracy of the model is better than the existing models.

The contribution of this paper is as follows:

1. This paper proposes achievement performance prediction model based on Deep Neural Network of student similarity, which can fully learn the similarity between the attributes of both students and students, and use the historical performance and similarity of similar students to predict student performance. At the same time, the model can make up for the shortcomings of a single Deep Neural Network that treats all students the same, and improve the prediction ability. The added Dropout can effectively solve the problem of model overfitting. The model has good prediction performance on public datasets, and the accuracy is higher than existing models.

2. By analyzing the publicly available dataset, this paper finds that there are some student-to-student connections and that there is a consistency of grades, with a strong connection between students' final grades and their grades in the previous two semesters. The more similar a student's historical grades are, the higher the likelihood that their final grades will be of the same grade level.

The rest of the paper is summarized below. Section II describes the related work on student performance prediction and deep learning. Part III describes achievement performance prediction model based on deep neural network of student similarity. The fourth section gives the experimental results, evaluation metrics, and data set analysis results. The last part is a concluding comment.

2 Related Works

2.1 Prediction of Student Achievement

Students' performance could be defined with different measures: exam grade or range, course grade or range, GPA, retention or dropout rate, knowledge gain[5]. Student performance prediction is an important issue in smart education, where machine learning and deep learning are utilized to predict and improve student performance in academics. The researchers uncovered the factors affecting students' performance through various aspects of students' personal information, behavioral information, social information, course information, and school information in order to improve the model performance. In the existing research, student performance prediction through machine learning algorithms is the key area of researchers' focus. Aggarwal et al[6] used academic and non-academic factors to demonstrate the importance of using non-academic factors to predict student performance. Non-academic factors were found to have a significant effect on student performance by training machine learning models such as Random Forest, Logistic Regression, Decision Tree, Bagging, MLP and AdaBoost.

The application of deep learning in the field of student performance prediction is gradually receiving close attention from scholars. Yao et al[7] collected and constructed a real dataset of students' campus online behavior and achievement data, and proposed an end-to-end two-layer self-attentive network, which introduces a cascading self-attention mechanism to extract students' local online behavioral features for each day and global online behavioral features for a long period of time, respectively, and better solves the problem of long behavioral sequence modeling. Yu et al[8] used graph neural networks to represent students' performance in various subjects and utilized multilayer perceptrons to learn the intrinsic relationship between subject performances. Li et al[9] proposed an end-to-end deep learning model that automatically extracts features from students' heterogeneous behavioral data from multiple sources to predict academic performance. The key innovation of the model is the use of a long and short-term memory network to capture the inherent time-series features of each behavior and a two-dimensional convolutional network to extract the correlation features between different behaviors. Sun et al[10] learned the relationship between students and courses through neural collaborative filtering as a way to predict student course performance.

2.2 Deep Neural Networks

The Deep neural network is a type of unsupervised neural network that usually has three layers, input layer, output layer, and hidden layer, with full connectivity between the layers. It learns the output of the previous layer as the input of the next layer, performs a series of linear transformations through the activation function, and outputs the vectors to the next layer, passing the output vectors all the way through.

The application of Deep Neural Networks in the prediction of student performance has also gained the attention of some scholars. Aya Nabil et al[11] utilized Deep Neural Networks to predict student performance using multiple resampling methods for the data imbalance problem and compared them, which was higher than traditional machine learning models in terms of accuracy. Yang et al[12] performed feature screening by two methods of random forest and then predicted student performance by an improved model based on K-means and Deep Neural Networks.

3 Methodology

In order to effectively utilize student learning data to identifying students' possible academic risks early in the semester. In this paper, we propose achievement performance prediction model based on Deep Neural Network of student similarity. The model framework diagram is shown in Figure 1.

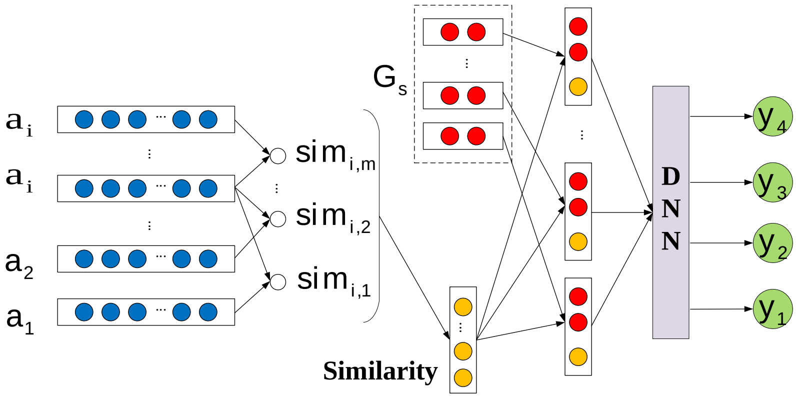

Fig. 1. Framework of Sim-DNN

Given a set of student features Z, Z consists of the student's attribute features $A = \{a_1, a_2, \dots, a_m\}^T$ and two historical grades $G = \{g_1, g_2, \dots, g_m\}^T$, m is the number of students. The student's predicted end-of-stage 3 grade is y, $y = \{y_1, y_2, \dots, y_c\}$ is the rating of the student's predicted grade. Student performance prediction is the prediction of the target student's final course grade at stage 3 based on the given characteristic attribute Z.

3.1 Student similarity

The analysis of the students' attribute profiles revealed that there was continuity in the students' performance across the three phases and that the more similar the students were, the more likely they were to be in the same grade where their final grades were located. There are three main ways of calculating similarity: Jaccard similarity coefficient, cosine similarity and Pearson correlation coefficient. The Jaccard similarity coefficient is an indicator of the similarity of two sets and is suitable for binary computation. Cosine similarity measures the similarity of two vectors by calculating the cosine of the angle between them. Pearson's correlation coefficient adds standardization of data to cosine similarity to reduce the effect of numerical paranoia. This model calculates the similarity of students by Pearson correlation coefficient as in (1).

Where $sim_{u,v}$ denotes the similarity of student u and student v, u_j and v_j denotes the jth attribute feature of student u and student v, $\bar{u}$ and $\bar{v}$ denotes the mean value of student attribute features, respectively. After obtaining the similarity between the target student and other students, the Top-K (K=30) most similar students are selected to form the similar students matrix S, as well as the similarity vector Sim and the similar students' historical performance matrix G_s based on the order of similarity from high to low. Weighted summation of the two history scores of similar students yields (2):

$$sim_{u,v} = \frac{\sum_{j=1}^{n}(u_j-\bar{u})(v_j-\bar{v})}{\sqrt{\sum_{j=1}^{n}(u_j-\bar{u})^2}\sqrt{\sum_{j=1}^{n}(v_j-\bar{v})^2}} \tag{1}$$

$$x_v = sim_{u,v}\left(\frac{g_{v1}+g_{v2}}{2}\right) \tag{2}$$

Where $sim_{u,v}$is the similarity between the target student u and student v, $g_{v1}$and $g_{v2}$are the two historical grades of student v.

3.2 Deep Neural Networks of student similarity

Deep Neural Networks enhance the expressive power of the model by expanding the hidden layers to multiple layers based on the perceptual machine. DNN mitigates the problem of locally optimal solutions through pre-training by using the output of the previous layer of the network as the input to the next layer of the network. In order to improve the generalization ability of the model and prevent the model from overfitting, this model adds Dropout to the DNN model and randomly selects some of the neurons in the hidden layer to make their output 0 for each training.

The model takes as input the historical grades and similarity-weighted summation of similar students of the target student as a means of predicting the target student's final grade at Key Stage Three. The DNN model is divided into three parts: input layer, hidden layer and output layer. x_1 tox_k be the input data to the model, which is output to the next layer through a nonlinear activation function in the input layer. The hidden layer's similarly taking the output of the previous layer as the input to this layer of the network and Through a series of nonlinear transformations, the which is the activation function of the hidden layer. The result is passed to the next layer. The layers are passed until the output of the final result y_i. If the output of the hidden layer h is x_h, the specific calculation process is as in (3):

$$x_h = f\left(\sum_{l=1}^{HL} w_l^h x_l^{h-1} + b_l^h\right) \tag{3}$$

Where f is the activation function of the hidden layer, HL is the number of neurons in the hidden layer h, $w_l^h$is the weight of the neuron, $x_l^{h-1}$is the output of the previous layer, and $b_l^h$is the bias term of this neuron. The hidden layer activation function is ReLU, which can effectively alleviate the problem of the gradient vanishing. The output of the last hidden layer is passed to the output layer, which outputs the final predicted value of the model, i.e., the probability of the target student's performance at each level, through the Sigmoid activation function.

4 Experiment

4.1 Datasets

This model student performance prediction model student performance[13] public education dataset is trained on two datasets of Portuguese and math performance. Two of the datasets include 30 student attribute features, as well as two historical grades, G1 and G2, and a final grade, G3. The student attribute profile describes the student's personal information, family situation, study habits, and social life. Table 1 illustrates the specifics of the student attributes.

Table 1. The description of dataset

ID	Attributes	Characteristics
1	School	CP=1,MS=0
2	Sex	F=1,M=0
…	…	…
30	Absences	Number of school absences(from 0 to 75)
31	G1	First historical achievement(From 0 to 20)
32	G2	Second historical achievement(From 0 to 20)
33	G3	Final grade(From 0 to 20)

This model analyzes whether there is a correlation between historical grades and final grades. The distribution of the three grades is very similar, with some continuity in student achievement in the absence of special factors. In the Portuguese dataset, 56.85% of the students' grades G1, G2, and G3 were at the same level, and this percentage was 61.34% in the math dataset.

The analysis of the datasets shows that there are instances of records with zero student achievements in the two public education datasets. Analyzing the attribute characteristics of these students, it was found that the students were normal in their studies and did not have all zero grades, and these students were considered as having missing grades due to other factors. Therefore, data with missing grades or grades of 0 were deleted and the remaining data with complete information were retained. After deletion, the math achievement dataset was 357 entries and the Portuguese dataset was 634 entries.

Student grades G1, G2, and G3 are numeric data, distributed among 0-20. The model categorizes grades into four levels A, B, C, and D. A grades are 16-20 points; B grades are 13-15 points; C grades are 10-12 points; and D grades are less than 10 points. A grade indicates that the student's performance is excellent, B grade indicates good, C grade indicates passing, and D grade indicates a failing.

4.2 Results

In the experiments, the dataset was randomly divided into a training set and a test set using 5-fold cross-validation. The network model is trained and tested based on the deep learning framework Pytorch, during training, Adam optimizer is used with Batch Size of 128 and Epochs of 100.This model uses 30 neurons as inputs in deep neural network and 4 neurons as outputs in the output layer with hidden layer dropout rate=0.5. Four evaluation metrics are used to assess the model of this model, which are accuracy, precision, recall and F1 Measure.

The proposed Sim-DNN is experimented with several classical machine learning models Nearest Neighbor Node Algorithm (KNN), Logistic Regression (LR), Random Forest (RF), Ridge Regression (Ridge), and Support Vector Machines (SVMs) on two publicly available educational datasets to validate the proposed methodology of this model.

From Tables 2 and 3, it can be seen that the proposed Sim-DNN has better performance on two publicly available educational datasets compared to several other classical machine learning prediction methods. The method in this paper achieves an accuracy of 89.36% on the math dataset and 91.32% on the Portuguese dataset, which is an improvement of 1.97% and 3.31% over the best performing of the benchmark methods, RF and SVM, respectively. In addition to this, the method in this paper on the other three classifiers precision, recall, and F1 Measure also achieved good performance, especially in F1 Measure the method is higher than other benchmark methods, reaching 92.66% and 94.84% on the two datasets respectively. The reason for the poor performance of the benchmark method on the dataset in the experiment may be that the traditional machine learning model inputs each attribute feature directly into the model as a categorization feature for learning and training, which does not take into account that students' learning is affected by different factors. They do not extract more effective information for student attribute features and ignore the similarity between students. And this model takes into account the fact that students' academic performance has continuity and there is a strong connection between historical grades and final grades. It predicts students' final grades through the similarity between students and achieves better prediction results, and the experimental results also prove its effectiveness.

To further validate the effectiveness of this paper, ablation experiments of Sim-DNN proposed in this paper are conducted on two publicly available educational datasets. This section compares the approaches of Deep Neural Networks and Deep Neural Networks based on student similarity, and the experimental results on the dataset are shown in Table 4. DNN refers to the standard DNN model.

Through experiments, it can be found that compared to the generalized DNN model, Sim-DNN proposed in this paper performs better in terms of accuracy and F1 Measure, which is 1.13% higher in the math dataset and 0.74% higher in the F1 Measure. These indicators reached 5.99% and 3.93% in the Portuguese dataset. The experimental results validate the effectiveness of the methodology of this paper to take into account the similarity of the students.

Table 2. Prediction results on math

Method	accuracy	Precision	Recall	F1-Measure
LR	74.78	74.71	89.59	81.48
SVM	78.43	77.36	92.34	84.19
KNN	82.07	80.00	95.07	86.88
Ridge	83.47	81.51	95.57	87.98
RF	87.39	84.53	98.24	90.87
CF-CNN	89.36	90.57	94.86	92.66

Table 3. Prediction results on portuguese

Method	accuracy	Precision	Recall	F1-Measure
LR	74.37	74.68	94.47	83.42
SVM	88.01	91.43	94.53	92.96
KNN	82.02	83.61	95.03	88.95
Ridge	83.43	82.51	98.05	89.61
RF	85.80	85.24	98.11	91.22
CF-CNN	91.32	91.98	97.86	94.84

Table 4. Ablation experiment on dataset

Data	Method	accuracy	Precision	Recall	F1-Measure
Math	DNN	88.23	90.19	93.72	91.92
	CF-DNN	89.36	90.57	94.86	92.66
Portuguese	DNN	85.33	84.69	981	90.91
	CF-DNN	91.32	91.98	97.86	94.84

4.3 Analysis of Factors Affecting Student Achievement

In order to further explore the factors influencing students' performance and to identify possible risks of failure in students' academics in a timely manner, we analyzes the factors influencing different G3 performance levels.

Students with an A grade category usually have parents who are well educated and have good family relationships. Fifty-five percent of the 40 students who performed with an A grade in math had mothers with higher education, and this percentage was 47.6% in the Portuguese dataset. At the same time, the family atmosphere was an important factor in the good performance of these grades. Only a few of the students who achieved an A achievement grade in the two education datasets perceived family relationships as poor, accounting for 7.5% and 3.6% of the total number of A grades, respectively.

Students with G3 grades of B had a variety of factors influencing their performance, closely related to the number of hours of study per week, whether they participated in activities and their mother's educational level. In the Portuguese dataset, 77.3% of the students maintained more than two hours of study per week, with 10 even studying more than 10 hours per week. 101 students chose to participate in activities to enrich their academic life and 35.1% of the students' mothers were highly educated. In the math dataset, only a quarter of the students study less than two hours per week; and more than half of the students choose to participate in activities; the proportion of mothers of students with higher education reaches 45%. It can be seen that most of the students who achieve good grades invest a lot of time and effort in their studies and also try to improve themselves by choosing to participate in different activities.

The number of C grades is the largest category in the overall student population, with aspects such as hours after school, mother's educational experience, physical health and hours of study per week having an impact on grades. In the math dataset, only 5.2% of the students had less time for after-school activities; close to one-fifth had mothers with only primary education; while close to one-third of them were in poor health; and close to one-quarter of the students stayed within two hours of studying per week. And these are 7.3.%, 26.7%, 20.5% and 38.4% of the Portuguese dataset, respectively.

Students who did not pass their academics were from a larger percentage of the rural population than the other achievement levels, and most of the students had more time off from school, and less than 30% of the students were in poor physical condition.

In conclusion, the educational status of parents and the amicability of family relationships have a greater impact on the academic performance of students. Families are the first to be responsible for a student's education, and a good family atmosphere is more likely to promote higher academic achievement. More educated parents are able to provide students with more academic counseling and more advanced educational concepts. The time invested in studying had a greater impact on students with achievement grades B and C, and a lesser impact on students with achievement grades A, suggesting that the number of hours spent studying is more important for middle-achieving students and that a certain amount of study investment is necessary. Poor physical health of students has increased among the low achievers, which can be considered as physical condition affecting the academic performance of students, and having a good body will enable them to engage in their studies in a better condition. Figure 2-5 shows the distribution of different levels.

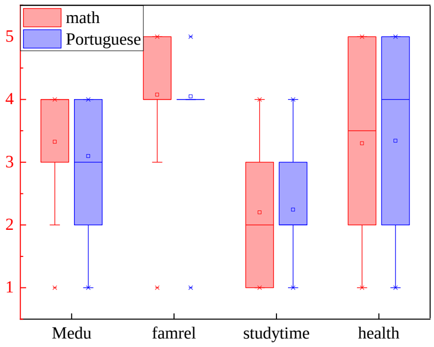

Fig. 2. A-grade distribution

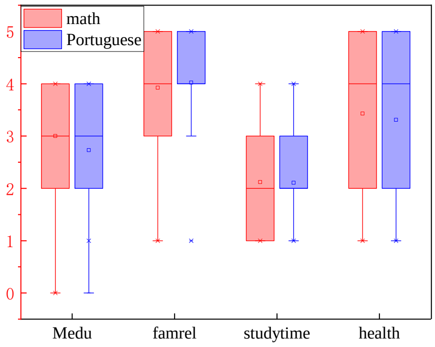

Fig. 3. B-grade distribution

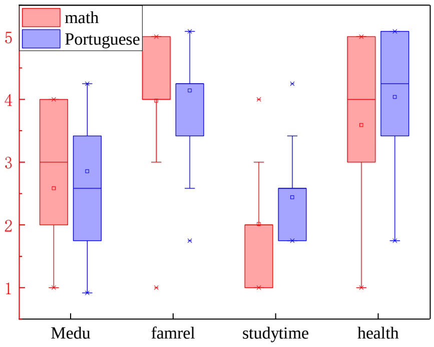

Fig. 4. C-grade distribution

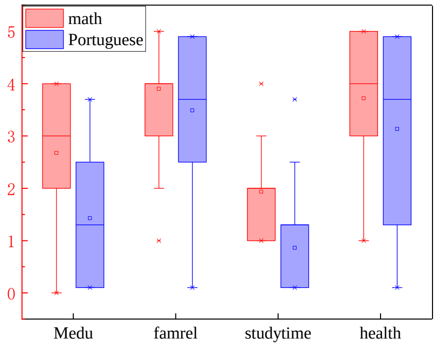

Fig. 5. D-grade distribution

5 Conclusions

Prediction of students' academic performance is one of the more popular areas of smart education in recent years and one of the most important means of identifying whether a student is at risk of failing academically. This paper collects two publicly available educational datasets based on math scores and Portuguese scores. Two publicly available educational datasets based on math scores and Portuguese language scores are collected. Student academic performance is predicted by means of Deep Neural Network of student similarity (Sim-DNN), which takes into account the connections between students in order to provide better guidance for the student learning. Student academic performance is predicted by Deep Neural Network of student similarity, taking into account the connections between students for better guidance in student learning. First, we find similar students by calculating the similarity between students. Then the historical scores of similar students and student similarity are utilized to predict students' academic performance through Deep Neural Networks. Considering the possible overfitting problem of the model, we add Dropout part to Deep Neural Network to improve the generalization ability of the model. The prediction model proposed in this paper achieves good performance in both public education datasets, which is higher than other benchmark methods, and fully proves the effectiveness of this paper's model in student achievement prediction.

This paper also visualizes and analyzes student achievement data. There is some continuity in student achievement without the influence of special factors. Factors affecting academic performance vary across achievement levels, but the mother's education and family atmosphere are very important factors. Mother's education is very important for students and even decisive for some of them. At the same time, a certain amount of time investment in learning is also necessary, and this factor has a greater impact on students with moderate performance. Finally, a healthy body is fundamental to learning; a good body can have a good state of learning, learning efficiency will be higher.

In future research, it is possible to consider the issue that different students have different influencing factors, and to take into account the issue of student individuality to provide more personalized guidance for student learning. Second, there is no standardized dataset for the prediction of students' academic performance, and it is hoped that more information on students' data can be collected, including performance in online and offline courses and students' performance in school, to provide more basis for the study. Meanwhile, for the problem of imbalance in most of the datasets, we can explore its effect on the prediction results.

References

[1] Xiao, W., Ji, P., & Hu, J. (2022)A survey on educational data mining methods used for predicting students' performance. Engineering Reports, 4(5): e12482. https://doi.org/10.1002/eng2.12482

[2] Cabero-Almenara, J., Llorente-Cejudo, C., & Martinez-Roig, R. (2022)The use of mixed, augmented and virtual reality in history of art teaching: A case study. Applied System Innovation, 5(3):44. https://doi. org/10.3390/asi5030044

[3] Francis B K,Babu S S. (2019)Predicting academic performance of students using a hybrid data mining approach.Journal of Medical Systems,43(6):1-15. https://doi.org/10.1007/s10916-019-1295-4

[4] Ghorbani,R., Ghousi,R.Comparing. (2020)Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. IEEE Access, 8:67899–67911. DOI: 10.1109/ACCESS. 2020.2986809

[5] Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., ... & Liao, S. N. (2018)Predicting academic performance: a systematic literature review. In: Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education. Cyprus. pp. 175-199.

[6] Aggarwal, D., Mittal, S., & Bali, V. (2021)Significance of non-academic parameters for predicting student performance using ensemble learning techniques. International Journal of System Dynamics Applications (IJSDA),10(3):38-49. DOI: 10.4018/IJSDA.2021070103

[7] Yao L, Cui C., Ma L. (2022) Campus Online Behavior-aware Student Performance Prediction. Journal of Computer Research and Development,59(8): 1770-1781. https://kns.cnki.net/kcms/detail/ 11.1777. TP. 20 220331.1521.010.html.

[8] Yu Y., Fan J., Xian Y.. (2022)Graph Neural Network for Senior High Student's Grade Prediction. Applied sciences, 12:3881. https://doi.org/ 10.3390/app12083881

[9] Li, X., Zhang, Y., Cheng, H., Li, M., & Yin, B. (2022)Student achievement prediction using deep neural network from multi-source campus data. Complex & Intelligent Systems, 8(6):5143-5156. https://doi.org/10.1007/s40747-022-00731-8

[10] Sun, H., Yin, C., Chen, H., Qiao, L., Ouyang, Y., & David, B. (2019) A student's performance prediction method based on neural collaborative filtering. In:2019 IEEE International Conference on Engineering, Technology and Education (TALE). pp. 1-8. DOI:10.1109/TALE48000. 2019.9225924

[11] Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021)Prediction of students' academic performance based on courses' grades using deep neural networks. IEEE Access, 9: 140731-140746. DOI:10.1109/ ACCESS.2021.3119596

[12] ang, X., Zhang, H., Chen, R., Li, S., Zhang, N., Wang, B., & Wang, X. (2022)Research on forecasting of student grade based on adaptive k-means and deep neural network. Wireless Communications and Mobile Computing, 2022. https://doi.org/10.1155/2022/5454158

[13] Cortez P. Silva A M G. (2008)Using data mining to predict secondary school student performance.In:Proc of the 15th Conf on European Concurrent Engineering. Ostend, Belgium: EUR()SIS, 5-12. https://hdl.handle.net/1822/8024