Research on Predicting Students' Performance Based on Machine Learning

Liu Ruochen[a], Mei Wenjuan[b], and Liu Jun[c,*]

School of Management Science and Engineering Nanjing University of Finance and Economics, No. 3 wenyuan road, Nanjing, China

[a]965703636@qq.com, [b]1563477026@qq.com, [c]9120031038@nufe.edu.cn

*corresponding author

Keywords: Teaching quality, support vector regression, decision tree

Abstract: Machine learning is one of the most core and hot technology of artificial intelligence at present. It can automatically identify patterns and discover rules based on a large amount of data, predict students' learning performance, and provide possibilities for more reasonable teaching evaluation and personalized learning. Taking the final mathematics scores of students in two Portuguese schools in the medium education as an example, this paper analyzes the characteristics of students' stage scores, personal personality, social relations and daily performance. After dimensionality reduction and other preprocessing of data sets by PCA and other methods, the final mathematics scores of students in one academic year are classified and predicted by SVM and decision tree algorithm respectively, and relevant factors affecting students' scores were analyzed. Finally, it concludes that schools can focus on students' family status, bad habit and ordinary grades to enable students to perform better.

1. Introduction

Under the influence and penetration of the wave of artificial intelligence research, education field is changing. On the one hand, with the support of mobile data collection tools and online collaboration platforms, the need of collecting and storing a large amount of data has been basically met. Such as Experience API (xAPI) is a technical specification for storing and accessing learning experiences [1]. At the same time, with the progress of technology, the teaching management information system of colleges and universities has been constantly improved, and the data of students' semester grades, classroom performance, social relations and other data can be obtained relatively easily and completely, thus providing a basis for the comprehensive analysis of student data. On the other hand, machine learning, as the most core and popular technology in the field of artificial intelligence, can automatically identify patterns and discover rules based on a large number of data, predict students' learning performance, and provide possibilities for more reasonable teaching evaluation and personalized learning [2]. Education evaluation refers to the process of judging the value of education on the basis of systematically, scientifically and comprehensively collecting, sorting, processing and analyzing education information. From a personal perspective, education evaluation aims to understand the development of students, objectively summarize the learning of students and evaluate the teaching quality of teachers. From a macro perspective, the purpose of evaluation is to promote education reform and improve the quality of education in the whole country [3].

In terms of student performance prediction, Bravo Agapito et al. used C4.5 decision tree rules to conduct a discriminant analysis of poor learning performance in online learning systems. Kabakchieva obtained a model that can predict students' final grades in 2013 by analyzing the characteristics of college students' personality, graduation school and daily behaviors [4,5]. In China, Hu Yunan et al., from Fudan University, put forward the collection, analysis and quality assessment scheme of learner model machine learning behavior on SCORM standardized online teaching management platform [6-7]. Bai Xuemei (2016) mainly studied the problem of using gradient descent method in machine learning to evaluate teachers' teaching work from the perspective of

  DOI: 10.25236/icbdai.2018.007

distinguishing teachers' performance in various work indicators [8]. These literatures rarely analyze the education evaluation from the perspectives of students' past performance, personal personality, social characteristics and school performance.

This paper will establish a model based on 32 different characteristics of students in four aspects and realize machine learning through Python to analyze students' relevant data. After preprocessing the data by means of principal component analysis and other methods, students' academic performance in this semester was classified and regression predicted, and the predicted results were evaluated and analyzed, so as to provide reference for teaching quality evaluation and improvement.

2. Data Collection and Preprocessing

2.1 Data Preparation

This paper referred to Paulo Cortez et al.'s literature and selected math-related data sets of students in two Portuguese schools in medium education [9]. Data were collected through school reports and questionnaires. After examination and testing, 395 records were finally obtained. Data attributes include students' grades at the stage, personal personality, social relations and daily performance. Part of the detailed attributes and descriptions are shown in Table 1. This paper selects the final score of mathematics as the prediction object. Students' grades are assessed in three stages throughout the school year, and the final assessment (G3 in Table 1) corresponds to the final grade.

Table 1: Description of data set characteristics.

feature	value range	Introduction of features
sex	'F' means female; 'M' means male;	student's sex
address	'U' means urban; 'R' means rural;	student's home address type
Medu and Fedu	numeric: 0: none; 1: base education(before 4th grade); 2: 5th to 9th grade; 3: secondary education or 4: higher education;	mother and father’s education
Mjob and Fjob	'teacher'; 'health' means care related; civil 'services' means civil services such as administrative or police; 'at home' or 'other'	mother and father’s job
studytime	numeric: 1: <15 min; 2: 15 to 30 min; 3: 30 min to 1 hour; 4: >1 hour;	weekly study time
failures	numeric: 1,2,3 or 4 means more;	number of past class failures
famrel	numeric: from 1 - very bad to 5 – excellent;	quality of family relationships
freetime	numeric: from 1 - very low to 5 - very high;	free time after school
Dalc and Walc	numeric: from 1 - very low to 5 - very high;	amount of workday and weekend alcohol consumption
absences	numeric: from 0 to 93;	number of absence from school
G1 and G2	numeric: from 0 to 20;	first and second period grade and
G3	numeric: from 0 to 20;	final grade

To facilitate calculation, the category label is first converted into a number variable, and then the missing value is searched for, and the average value is used to fill in the missing value. Pandas' describe function is used for descriptive statistics. According to the descriptive statistical results, we can see the general distribution of the samples, in which the proportion of men and women is balanced, and most students live in urban areas and are accompanied by their parents. Since most of the students with a final score of 0 did not participate in the examination, records with a final score of 0 were deleted in the data preprocessing stage, and the number of samples processed was 357. Figure 1 shows the distribution of different grades after processing.

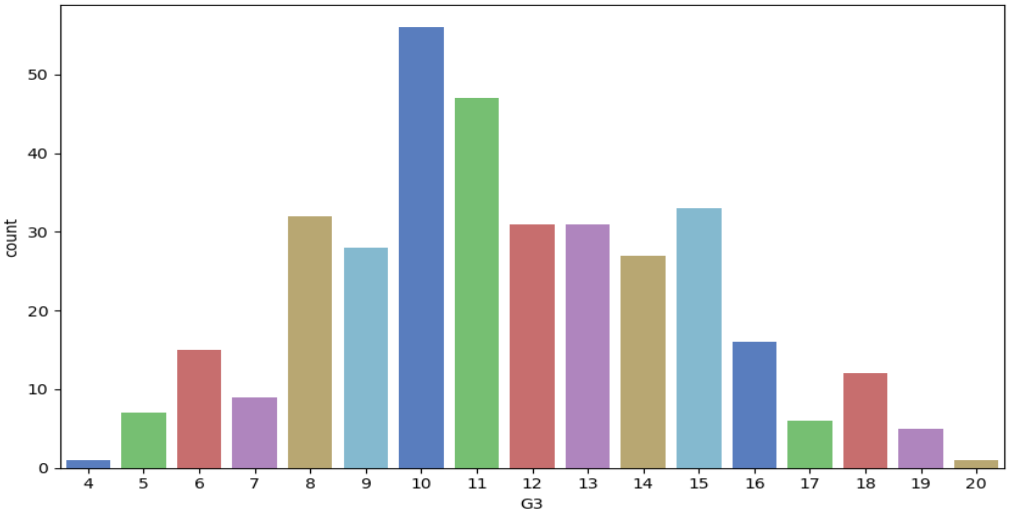

Figure 1 The histogram of different grades after processing.

Classify the grades and add classification labels. When the score is less than 10, the grade is "low"; when the score is between 11 and 15, the grade is "medium"; when the score is greater than 15, the grade is "high". The histogram after classification is shown in Figure 2.

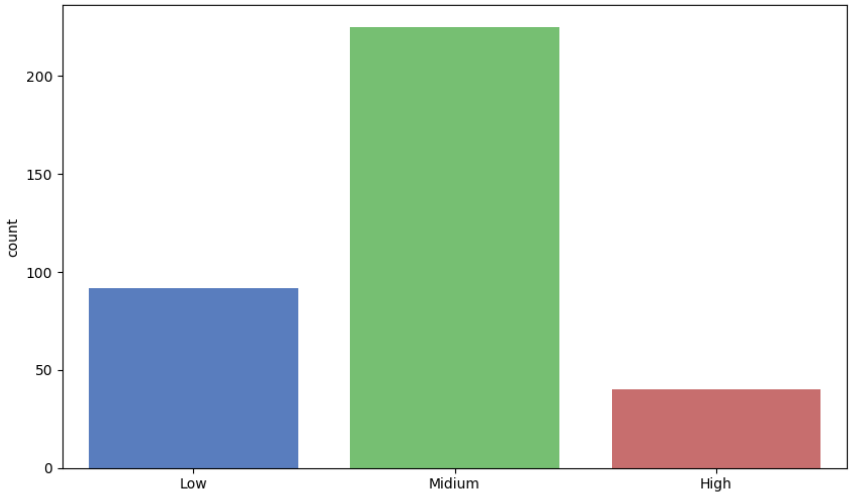

Figure 2 Histogram of final grade distribution.

2.2 Principal Components Analysis (PCA)

PCA is a statistical analysis technology for studying multivariate problems[10]. The basic way of PCA is to convert the initial sample data of a given multidimensional degree into another set of linearly unrelated data through linear transformation according to certain mathematical transformation. The converted data are sorted by the variance value from large to small. Under the condition that the total variance of the initial sample data remains unchanged, the linear transformation is carried out to maximize the variance value of the first dimensional data obtained after the transformation, that is, the first principal component; Similarly, the second dimension data is called the second principal component, whose variance value is only less than the first dimension, and is linearly unrelated to the first dimension data. By the way, the lower the variance of the last principal component is, the smaller its contribution to the computing task will be.

The principal components obtained by PCA have the following basic relations with the original data: The amount of principal components after conversion is greatly reduced, and the principal

component is obtained by linear combination of pre-conversion data, so it also contains the content of pre-conversion data. The principal components are independent and orthogonal, and their information does not overlap.

PCA can sort the results of principal component analysis by importance and select only the components that can represent the meaning of most indicators according to the user's needs, so as to simplify the model or compress data. In addition, PCA has no parameter limitation, no need to set artificial parameters, and has strong practicability.

In this paper, PCA is used to extract the features of information related to students. Through PCA, it can be observed that the sum of variance of the first four principal components accounts for 90.639% (>85%) of the total variance, which can represent the main information of the data source. When using SVM algorithm for regression prediction, the data after dimensionality reduction will be input to train the model.

3. Model and Training

The purpose of prediction is to develop a model to infer the possible values of one feature of a sample based on other features of the sample, in short, the process of extrapolating an unknown event from a known event. In the application of education, the commonly used prediction methods are Classification and Regression, which are generally used to predict student performance and detect student behavior. In this paper, support vector regression is used to predict students' math scores in this academic year, and decision tree algorithm is used to classify students' math grades.

3.1 Regression

3.1.1 Support Vector Regression (SVR)

SVR is a common regression method in machine learning. For general regression problems, the value of parameter is determined by training the training dataset, to obtain an $f(\boldsymbol{x}) = \boldsymbol{\omega} \cdot \boldsymbol{x} + b$, and the value of $f(\boldsymbol{x})$ is as close as possible to y. In this model, the loss is zero only if $f(\boldsymbol{x})$ and y are completely identical, and SVR assumes that the maximum deviation between $f(\boldsymbol{x})$ and y that we can tolerate is ε. Losses are calculated if and only if the absolute difference between $f(\boldsymbol{x})$ and y is larger than ε. At this time, it is equivalent to constructing an interval band with a width of 2ε with $f(\boldsymbol{x})$ as the center. If the training sample falls into this interval band, it is considered to be correctly predicted (the degree of relaxation on both sides of the interval band may be different).

3.1.2 Kernel Function

When the support vector machine is linearly separable or almost linearly separable, it can directly establish the hyperplane in the original space as the classification plane. However, most problems in practical applications are complex and nonlinear, so it is necessary to seek complex hyperplanes as classification planes. Support vector machines (SVM) establish a hyperplane of classification by using the method of dealing with linear problems in another high-dimensional space, thus implicitly establishing a hyperplane in the original space. The support vector machine (SVM) method avoids the calculation of high-dimensional space, and does not explicitly transform, but only performs the inner product operation between training samples, which is implemented by pre-defined kernel functions. Support vector machine uses kernel function to map the problem of linear indivisibility in linear space to the higher-dimensional nonlinear space, and thus becomes a linear problem, In this way, the complex computation problem of linear indivisibility is solved.

In this paper, when using support vector regression to predict students' performance, regressions based on rbf kernel, linear kernel and polynomial kernel are respectively adopted. Training sets after dimensionality reduction of PCA are taken as samples for learning, and regression effects of trainers using different kernel functions are compared. The comparison between students' actual score and predicted score is shown in Figure 3.

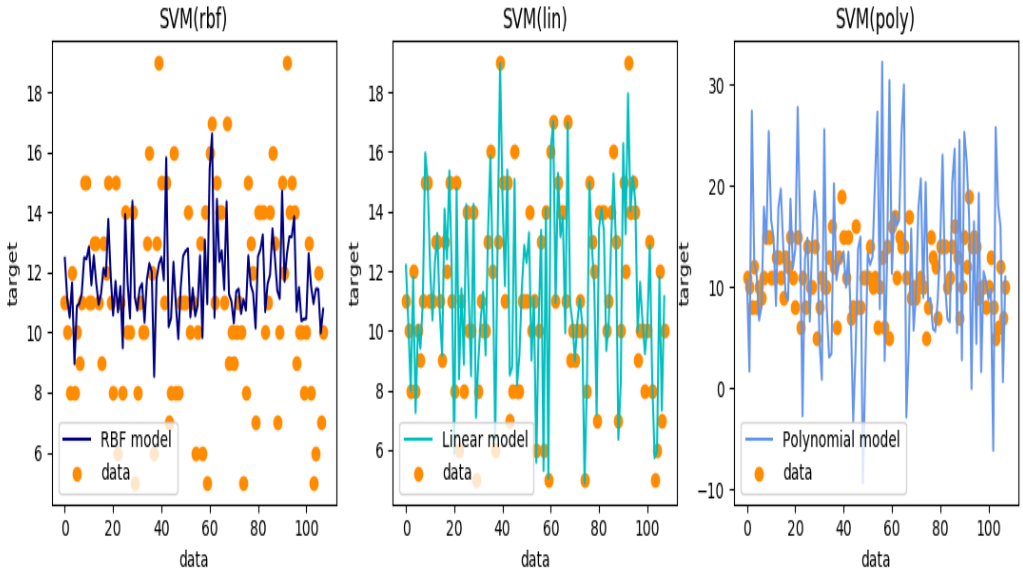

Figure 3 SVM prediction results of different kernel functions.

3.2 Classification

3.2.1 Decision Tree

As a common supervised classification algorithm, decision tree inferences the branching mode of the tree from a set of given tuples which are unordered and irregular [11]. For each problem or event, the decision making may cause two or more subproblems, and produce different results. The decision tree adopts the top-down recursion method, and for each decision branch, it can be recursed in the same way. The whole decision-making context is similar to the structure of the tree, hence the name decision tree. In addition to the leaf node, each other node means a test of a set of sample features, and the branch stands for the test output. The internal nodes conduct attribute value comparison and branch down from this node according to different attribute values. Leaf nodes are classes to be divided.

The algorithm idea of decision tree can be divided into three steps: selecting split feature, generating decision tree and adjusting decision tree. Firstly, the index values of different features are calculated according to an index (relative entropy, relative entropy ratio or gini coefficient), and take the best characteristic to be the split node. After that, gradually downward, repeatedly split feature selection to generate child nodes, and stop the growth of the decision tree when the sample set is indivisible. The fully expanded decision tree obtained from the first two steps is generally easy to be overfitted, so it needs to be pruned to a certain extent (to reduce the depth of the decision tree or the number of child nodes).

3.2.2 Grid Search with Cross Validation

Cross validation is often used to adjust the parameters of trainers in machine learning. The principle of cross-validation is to make use of the total sample for many times. Firstly, the obtained sample data are segmented, and then the segmented data are respectively combined into different training sample and testing sample. The model is trained on the training sample, and the prediction effect of the learning is evaluated on testing sample. Therefore, through the combination of different samples, we can obtain more than one set of training dataset and testing dataset.

Grid search method is to select the best performing parameter as the final result in all candidate parameter selection by loop traversal, trying every possibility. The final performance of grid search method has a great relationship with the partition result of the initial data, so to solve this problem, cross-validation is usually adopted to reduce the chance. The combination of the two as a parameter evaluation method, called the use of cross-validation grid search method. In python, parameters are adjusted using the GridSearchCV class of the sklearn package.

In this paper, the data are divided into 10 parts for cross validation, and the evaluation criteria and maximum depth are selected as the parameters to be optimized. The evaluation criteria include gini coefficient and information entropy, and the maximum depth ranges from 3 to 12. Finally, the gini coefficient was selected as the evaluation standard and the maximum depth was 3.The resulting decision tree is shown in the figure 4.

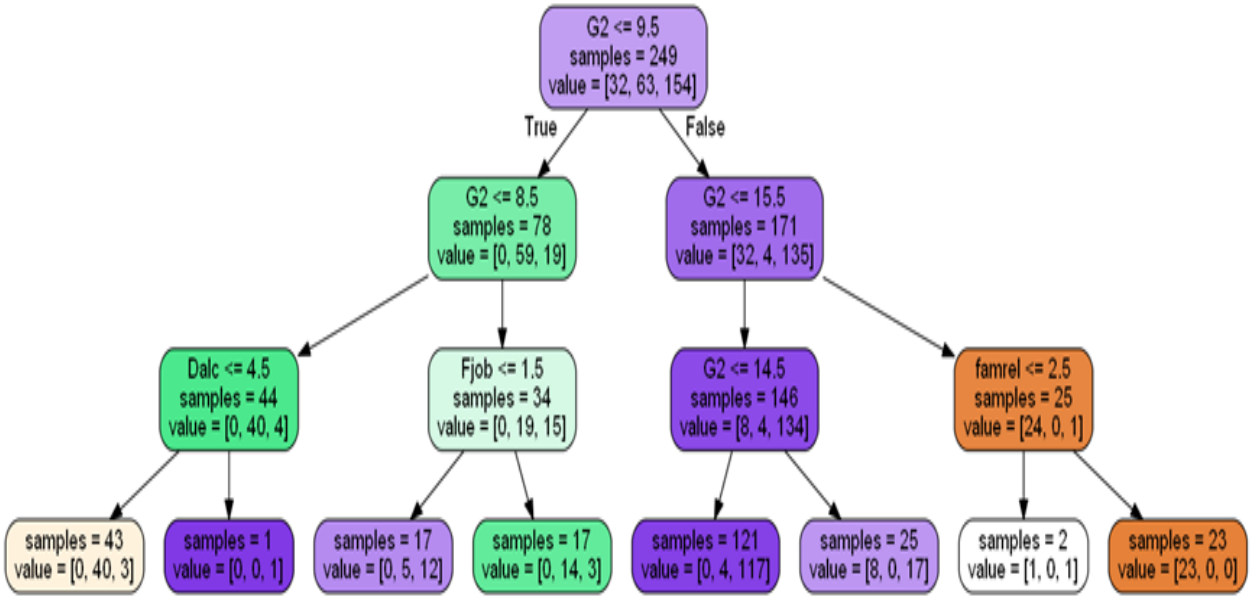

Figure 4 Students' performance classification decision tree.

4. Results Evaluation and Analysis

4.1 Model Evaluation

4.1.1 Regression Effect Evaluation

For the regression model, Mean Absolute Error (MAE), Mean squared error (MSE) and r2_score are adopted to compare and evaluate the regression effect. MAE compares and calculates the predicted results with the real data and determines their proximity. The smaller the value is, the better the fitting effect is. MSE calculates the mean square sum of errors between fitting data and original data corresponding to sample points. When its value is small, the fitting effect is better. R2_score, the judgment coefficient, means the variance score of the regression model. This index indicates the proportion of the total sum of squares of dispersion to the value of the sum of squares of regression. the smaller the value, the worse the effect.

The regression effect score is shown in table 2.The results show that the effect of linear kernel function is better than the other two kernel functions under different evaluation criteria. Therefore, the SVR of linear kernel function can be used to obtain the ideal effect in the prediction of students' academic year score.

Table 2 The comparison table of regression effect evaluation.

	SVM(rbf)	**SVM(lin)**	**SVM(poly)**
MAE	1.75646772683	0.598769855579	6.81551578066
MSE	5.55758364713	0.64903985812	76.9928722689
r2_score	0.441189834227	0.934739611	-6.74156584006

4.1.2 Classification Effect Evaluation

Confusion Matrix is a visualization tool used in machine learning especially for supervised learning. It summarizes the records in the data set in the form of Matrix according to the real category and the classification judgment made by the classification model. Each of its columns represents the predicted value, the row means the correct category, and the element in the matrix means the number

of categories that are taken to be another category.

This paper uses the decision tree to classify the grades of students, and the confusion matrix obtained according to the result is shown in figure 4. Among them, "0" represents the level of "high", "1" represents the level of "low", and "2" represents the level of "medium".

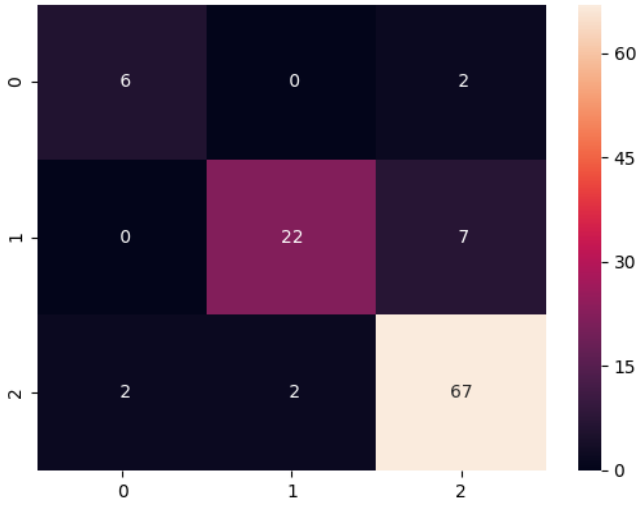

Figure 5 Decision tree classification confusion matrix.

Many indicators can be obtained from the confusion matrix, and the classification effect can be intuitively observed. The commonly used ones are accuracy, precision, recall, as well as F1 score. Precision is the amount of right predictions divided by the total amount of predictions. Recall represents recall rate, which is the proportion of the predicted correct number to the actual amount of samples. The recall rate is for the sample, indicating what percentage of the sample in a certain category was predicted correctly. It describes coverage and represents the ability of a model to find samples.F1-score is a composite index of precision rate and recall rate. In fact, it is the average after adjusting the recall rate and precision rate. These indicators can help analyze classification results from different perspectives.

This paper uses the decision tree to classify the grades of students with an accuracy of 0.8796. The evaluation results are shown in table 3.It can be found that the values of each index are more than 0.75, which can be applied to the actual grade classification.

Table 3 The evaluation table of decision tree classification effect.

	precision	**recall**	**f1-score**	**support**
low	0.92	0.76	0.83	29
medium	0.88	0.94	0.91	71
high	0.75	0.75	0.75	8

4.2 Result Analysis

It can be concluded from the observation of figure 4 that the students' performance in the second stage, the amount of alcohol consumed during the working day, the father's work and family relationship has a great impact on the students' final performance. Students who scored less than 9.5 in the second stage and who drank alcohol during the working day were most likely to have a lower final grade. Therefore, education institutions should strengthen the management of students' habits at ordinary times and prohibit students from drinking a lot of alcohol during non-holidays. At the same time, students whose grades in the second stage are between 8.5 and 9.5 and whose fathers work as non-teachers or health care types are likely to get low scores in the final stage, while students whose scores in the second stage are greater than 15.5 and whose family relationship is harmonious are very likely to get high scores in the end. So the school should care about the students' family situation and

give them the help they need. In addition, as can be seen from the decision tree diagram, the usual grades play a very important role in the prediction of students' final grades. Therefore, the school conducts more stage tests to keep an understanding of students' learning level and provide references for the development of corresponding teaching quality improvement activities.

5. Conclusion

Taking the final mathematics scores of students in two Portuguese schools in the medium education as an example, this paper analyzes the characteristics of students' stage scores, personal personality, social relations and daily performance. After dimensionality reduction and other preprocessing of data sets by PCA and other methods, the final mathematics scores of students in one academic year are classified and predicted by SVM and decision tree algorithm respectively, and relevant factors affecting students' scores were analyzed, which provided reference for education institutions to carry out corresponding teaching quality improvement activities.

Three conclusions are drawn from the experiment. Firstly, the school can enhance students' attention to their usual grades and grasp the changes in their grades, so as to take corresponding measures to maintain or improve their abilities. Second, pay attention to the family status of students, and pay attention to students whose fathers are not teachers or health care workers. Third, students' daily performance should be controlled to a certain extent, and students are prohibited from drinking a lot of alcohol during the working day.

The shortcoming of this paper is that it does not analyze the influence of different features on students' performance. For example, among the features related to daily performance, which factors have a greater impact on students' performance; follow-up research can further analyze the influence of different features on students' performance prediction.

Acknowledgement

This paper was supported by the Jiangsu province higher education education reform research project, 2017, project number：2017JSJG218 and Postgraduate Research & Practice Innovation Program of Jiangsu Province, 2018, project number: KYCX18_1325.

References

[1] Chiang, C. F., Tseng, H. C., Chiang, C. C. and Hung, J. L. (2015) A case study on learning analytics using Experience API. In Society for Information Technology and Teacher Education International Conference (pp. 2273-2278). Association for the Advancement of Computing in Education (AACE).

[2] Wang D. and Wang H. (1999) Pedagogy. Beijing: people's education press, 290-293.

[3] Yu M. (2017) Education application and innovation exploration of machine learning from the perspective of artificial intelligence. Journal of remote education, (1), 11-21.

[4] Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. Cybernetics and Information Technologies, 13(1).

[5] Kabakchieva, D., Stefanova, K., and Kisimov, V. (2011). Analyzing university data for determining student profiles and predicting performance. In 4th international conference on educational data mining (Eindhoven, the Netherlands).

[6] Li B. (2016) Research on teaching evaluation based on big data. Modern education technology, 26(6), 5-12.

[7] Zhu H. (2009) Research on teaching quality evaluation based on SVM multi-classification. Shandong normal university.

[8] Bai X. (2016) Research on teaching evaluation system based on machine learning. Education

teaching BBS, (15), 173-174.

[9] Cortez P. and Silva A. (2008) Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008). 5-12, Porto, Portugal, April, EUROSIS.

[10] Zhou Z. (2016) Machine learning. Beijing: tsinghua university press, 121-145.

[11] Strecht, P., Cruz, L., Soares, C., Merdes-Moreria, J. and Abren, R. (2015). A comparative study of classification and regression algorithms for modelling students' academic performance. In 8th international conference on educational data mining, Madrid: Spain, 392-395.