J. COMBIN. MATH. COMBIN. COMPUT. 127a (2025) 7659--7676

Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on Multidimensional Data Calculation Analysis and Optimization Path of Teachers' Practical Teaching Cultivation Innovation Mechanisms under Digital Empowerment of Education

Lina Zhang[1, ✉]

[1] School of Life and Health, Zhengzhou Vocational and Technical College, Zhengzhou, Henan, 450000, China

ABSTRACT
With the development of big data and education informatization, education reform and talent cultivation mode are facing digital reform. In this paper, the important feature selection algorithm based on random forest is used to select the relevant features that affect the application effect of teachers' practice teaching cultivation and innovation mechanism, which lays the foundation for constructing the practice teaching data mining model based on LightGBM. Then the data processed by feature selection is preprocessed and standardized, and then the processed data is partitioned and the model is trained in turn to get the prediction results. The LightGBM-based practical teaching data mining model was compared with other classification models in different datasets, and the experimental results showed that the model in this paper has an advantage over other classification models in a number of evaluation indexes, with the highest accuracy rate of 13.07%, and the model data mining results accurately locate the open innovation experimental indexes that have a lower score of importance to students' development, and provide a good basis for the optimization of teaching paths and students' development. , which provides ideas for the optimization of teaching paths and the improvement of the impact of students' future development.

Keywords: LightGBM, multidimensional data computation, random forest, data mining

1. Introduction

The rapid development of digital technology and the rise of the information society, the digitalization of education has become one of the trends in global education. With the increasing demand for education from students, parents and the society, traditional education methods have been difficult to meet people's needs. In the traditional form of education, students' monotonous classroom learning

✉ Corresponding author.
E-mail address: zhanglina098210@163.com (L. Zhang).
Received 20 February 2024; Revised 15 April 2024; Accepted 20 November 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-424

and boring teaching methods are unable to stimulate students' interest in learning. At the same time, the solidified education concept hinders the improvement of teachers' teaching level and professional development [1-3]. While digital education combines the digital classroom with the traditional classroom, changes the teachers' teaching philosophy, improves teachers' digital literacy, combines the students' needs and the operability of teaching, and can better promote the development and innovation of education [4-5].

Education is the mainstay of a hundred-year plan. Teachers are the foundation of education. Teachers are the first resource for the development of education and an important cornerstone for the wealth and strength of the country, the revitalization of the nation and the happiness of the people. The external law of education usually lies in the active adjustment of its own talent training methods and training paths, prompting education to actively participate in maintaining the prosperity of the regional economy, for the high-quality development of education, the key lies in the improvement of the quality of technical and skilled personnel training, and the core of improving the quality of personnel training is to improve the level of the teaching staff, that is, high-quality teaching staff is the first resource and the key force to promote the development of education quality. That is to say, high-quality teaching force is the first resource and key force to promote the high-quality development of education [6-8]. Teacher team building is a soft factor that determines the quality of teaching and the improvement of the quality of technical skills training in schools. Teachers, as the main bearers of talent training in schools, are the center of education and teaching activities, and the overall level of the teacher team will directly constrain the development of students' knowledge and skills and vocational ability, which directly affects the quality of education [9-12]. Practice teaching is the key to improve the quality of education, and cultivating teachers' practice teaching ability is an important factor in practice teaching. Teachers' practical teaching cultivation, on the one hand, can improve students' learning interest and participation, and improve the effect and quality of teaching [13-14]. On the other hand, it can enhance teachers' professionalism and educational concepts, improve the quality of education and teaching, and at the same time promote teachers' professional development and innovation, and promote the long-term development of education [15-16]. Innovating education and teaching methods, strengthening practical teaching links, and forming new advantages in talent cultivation are the guiding guidelines for improving practical teaching [17]. In the context of digital education, it is important to explore the innovative mechanism of teachers' practical teaching cultivation and optimize its path to contribute to the realization of high-quality education.

This paper proposes an important feature selection algorithm based on random forest and a data mining model based on LightGBM. The important feature selection algorithm based on Random Forest carries out feature selection for the application effect of practical teaching cultivation innovation mechanism by comparing the contribution degree between each feature. The LightGBM algorithm in the data mining model of teachers' practice teaching cultivation innovation mechanism adopts the gradient-based unilateral sampling algorithm and the mutually exclusive feature bundling algorithm, which significantly improves the speed of teaching practice data mining. The gradient-based one-sided sampling algorithm selects sample points with larger gradients to improve the contribution of information gain. The mutually exclusive feature bundling algorithm, on the other hand, further improves the running speed of the multidimensional data mining model in the teaching practice data mining task by dimensionality reduction of effective features. The usability of the algorithms in this paper for teaching practice data mining and teaching path optimization is verified through comparative experiments and practical tests.

2. Practice teaching multidimensional data analysis methods

2.1. Algorithm for selecting important features of student behavior

2.1.1. Random Forests. Random Forest (RF) [18], which is sampling according to a randomized method, is able to build a decision forest that satisfies the demand, and inside this forest are multiple decision trees. Therefore Random Forest is an integrated classification algorithm consisting of several decision trees, which is based on the decision tree algorithm, and its core is how to use the idea of randomization to build the multiple decision trees of Random Forest. Each decision tree produces a prediction result, and as many decision trees as there are in the forest can produce as many results. By integrating each decision tree, taking the strengths and weaknesses and deciding on the final prediction output, it can avoid limitations and have a better generalization ability, so usually the random forest is better than the prediction made by the single classification model.

Among the integrated algorithms, Random Forest is not a complicated algorithmic idea, and it is a more common method in dealing with binary classification, multiple classification and regression tasks. And because each decision tree is independent of the parallel operation, which can save time and computational overhead, is called “integrated learning technology level of representative methods”. Of course, Random Forest has its shortcomings, Random Forest does not continuously output the results, and if the data is out of the range of the training set, it becomes difficult to predict, so it may lead to overfitting, but the problem of overfitting will not be further investigated in this paper. On the other hand, Random Forests have no control over their internal structure of operation, so there is no way to intervene in the process, and if multiple similar decision trees are generated, then the true results are likely to be masked.

2.1.2. Random Forest based algorithm for important feature selection. The idea of random forest algorithm used for feature selection is the degree of contribution of each behavioral feature in each tree in the random forest, and feature selection is carried out by comparing the size of the contribution between features, and the degree of contribution of the features needs to be measured by the feature selection evaluation criteria (feature evaluation function). The calculation method of feature importance score for feature selection of random forest algorithm is mainly through the out-of-bag data error rate and Gini index, which is usually selected according to the data distribution characteristics and the type of data, and the Gini index is unbiased as the evaluation of the feature importance when the variables are continuous type and uncorrelated with each other. It has been proved that the accuracy of the Gini index evaluation criterion is higher than that of the out-of-bag data error rate when the signal-to-noise of the data is regardless of low and high, and the Gini index evaluation criterion has higher stability. While the utilization of out-of-bag data error rate as the evaluation criterion is more widely applied, the accuracy of out-of-bag data error rate is prone to be affected to some extent when the categorical data are unbalanced. In summary, comparing the applicability and stability of the above two, it is relatively better to choose the Gini index, therefore, this paper chooses the Gini index as a measure of the effect of feature segmentation.

The advantages of using the important feature selection algorithm based on the random forest algorithm are: first, the random forest algorithm uses the BAGGING integration framework to combine multiple decision trees, which is conducive to improving the model generalization ability compared with the single decision tree algorithm. Second, compared with the boosting integration framework, the random forest algorithm represented by the bagging integration framework is implemented using parallelization, and each base learner training obtained is independent of each other and is highly efficient; third, the random forest can better deal with high-dimensional datasets, is not prone to overfitting phenomena, and has a greater tolerance for classification imbalance in the data samples. Next, the evaluation function of the important feature selection algorithm based on random forest is introduced, i.e., the Gini index evaluation function.

Now the importance score of the feature is denoted by VIM and the Gini value is denoted by G. Assuming that there are m features $x_1, x_2, x_3, \cdots x_m$, now the average change in the split impurity of all the decision tree nodes in the random forest is calculated for each feature x_i, i.e., the i rd Gini index score $VIM_j^{(Gini)}$. Then the Gini index can be expressed as:

$$G_m = \sum_{k=1}^{|K|} \sum_{k' \neq k} p_{mk} p_{mk'} = 1 - \sum_{k=1}^{|K|} p_{mk}^2 \tag{1}$$

where K denotes K categories and p_{ak} denotes the proportion of categories being k in node m.

Secondly, the importance of feature x_i in node m is:

$$VIM_{jm}^{(Gini)} = G_m - G_l - G_r \tag{2}$$

where, G_1 and G_r denote the Gini indices of the two new nodes after branching, respectively.

If, the nodes where feature x_i appears in decision tree j are set M, then the importance of x_i in the j th tree is:

$$VIM_{ji}^{(Gini)} = \sum_{m \in M} VIM_{im}^{(Gini)} \tag{3}$$

Assuming that there are n trees in the random forest, the Gini importance of feature x_i in the random forest can be expressed as:

$$VIM_i^{(Gini)} = \sum_{j=1}^{n} VIM_{ji}^{(Gini)} \tag{4}$$

Finally, the calculated importance scores were obtained by processing them in a normalized manner:

$$VIM_i = \frac{VIM_i}{\sum_{j=1}^{m} VIM_j} \tag{5}$$

In summary the importance of each feature can be calculated and each feature can be ranked.

2.2. LightGBM-based Data Mining for Practical Teaching and Learning

In the task of multidimensional data mining of teachers' practice teaching cultivation and innovation mechanism, the important feature selection algorithm based on random forest can be applied to select the important features in the data, and then combined with the LightGBM model to carry out efficient large-scale multidimensional data mining, in order to more efficiently and comprehensively realize the multidimensional data mining of the teachers' practice teaching cultivation and innovation mechanism, and establish a reasonable data support to explore the optimized paths of the practice teaching cultivation and innovation mechanism. Optimization path to establish reasonable data support.

2.2.1. LightGBM Principles. Before introducing the LightGBM [19] (Light Gradient Boosting Decision Tree) algorithm, the framework of Boosting algorithm is introduced, which boosts weak learners into strong classifiers. Traditional algorithms for Boosting include GBDT [20] and XGBoost, which have been researched in various fields.

GBDT is an integrated learning framework, i.e., using the boosted decision tree as the base and the previous generation residuals as the gradient, it continuously trains new decision trees and accumulates them as the final strong learner model. Compared with a single decision tree, GBDT has stronger nonlinear fitting ability and prediction accuracy. In the regression problem let the training set be $\{(x_i, y_i)\}^n i = 1$, where x_i is the i rd input feature and y_i is the target value. At the m th iteration step, the predicted value $y'i(m)$ of the model can be expressed as:

$$y'^{(m)} = y_i(0)\sum_{k=1}^{m} \gamma_k h_k(x_i) \tag{6}$$

where: $y_i(0)$ is the initial predicted value, usually taken as the mean of all target values; γ_k is the learning rate in step k; $h_k(x_i)$ is the decision tree fitted in step k.

The decision tree $h_k(x_i)$ can be expressed as:

$$h_k(x_i) = \sum_{j=1}^{J_k} c_k, j(x_i \in R_{k,j}) \tag{7}$$

where: j_k the number of leaf nodes of the decision tree in step k; c_k is the predicted value of the decision tree in step k at leaf node i; $R_{k,j}$ the feature region corresponding to the leaf nodes of the decision tree in step k.

Extreme Gradient Boosting Algorithm, also known as XGBoost algorithm, is one of the inherited algorithms of boosting algorithm. In contrast to GDBT, XGBoost chooses to use the first-order derivative together with the second-order derivative, and the regular term in the objective function selects the complexity of the tree model, thus reducing the overfitting situation.

The objective function of the algorithm is assumed to be:

$$L(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{8}$$

where $\Omega(f) = \gamma T + \lambda \| w \|^2 /2$, $L(\varnothing)$ are loss functions, usually convex, that measure the difference between the predicted value y_i and the true value y_i; and $\Omega(f\varnothing)$ is a regularization function that reduces model complexity and mitigates overfitting. The optimization objective is to minimize the objective function.

The loss function is obtained by performing a second order Taylor expansion at $\hat{y}_l(t)$:

$$\begin{aligned} L^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_l^{(t-1)} + f_t(x_i)) + \Omega(f_t) \\ &\approx \sum_{j=1}^{T}\left[gf_t(x_i) + \frac{1}{2} hf_t(x_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \end{aligned} \tag{9}$$

In the above equation, g_i is the first order derivative and h_i is the second order derivative.

Definition $I_j = \{i \mid q(x_i) = j\}$ is the set of samples on leaf node j, then:

$$L^{(t)} = \left[(\sum_{i∂I_j} g_i) w_j + \frac{1}{2} (\sum_{i∂I_j} h_i + \lambda) w_j^2 \right] + \gamma T \tag{10}$$

When the tree structure q is known and the leaf node weights w_j of the above equation have a closed-form solution, the objective function:

$$w^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{11}$$

$$L^{(t)}(q)=\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j}g_i\right)^2}{\sum_{i\in I_j}h_i+\lambda}+\gamma T \tag{12}$$

But the disadvantage of Boosting algorithm is that all the samples have to be scanned for points to select the best segmentation point for each training, which leads to time-consuming training. And LightGBM algorithm further improves on the drawbacks of GBDT.

Light gradient boosting, also known as LightGBM, is an improved open source algorithm with the same properties as Xgboost, although both belong to the gradient boosting tree, but usually runs faster, takes up less content, and solves the efficiency problem in the environment of large samples and high latitude data. The gradient-based one-sided sampling algorithm and the mutually exclusive feature bundling algorithm will be used in LightGBM.

The gradient-based one-sided sampling algorithm will select the sample points with larger gradient to contribute more information gain when calculating the information gain. Therefore, when down-sampling the samples, the sample points with large gradients are retained, while the sample points with small gradients are sampled proportionally and randomly as a way to improve the efficiency. And the main idea of the mutually exclusive feature bundling algorithm is to reduce the number of effective features in order to achieve the purpose of improving efficiency by reducing the feature dimension.

LightGBM mainly has the following three important features:

1) Leaf-wise leaf growth strategy with depth restriction. LightGBM measures the use of leaf-wise strategy to grow the tree, it will select the node with the largest classification gain from all the leaves for splitting, and repeatedly perform the cycle and set the maximum depth, so as to avoid the appearance of overfitting.

2) Directly support feature category features. The LightGBM algorithm directly supports category features, which many machine learning algorithms cannot directly support. LightGBM is optimized for the support of the category, so that it can be directly input, without further conversion into a multi-dimensional one-hot coding form, which saves time and reduces space.

3) Support for efficient parallelism. LightGBM has the feature of supporting feature parallelism as well as data parallelism. Feature parallelism means that different machines search for the optimal segmentation points in different feature sets and synchronize the optimal segmentation points on different machines. Data parallelism means that different machines construct local histograms, merge them globally, and find the optimal segmentation points on the final merged histogram.

2.2.2. Classifier construction based on LightGBM. LightGBM is a further improvement of GBDT algorithm and XGBoost algorithm, which not only improves the training speed, but also enhances the prediction ability of the model. Combined with the above theory, this chapter proposes a multidimensional data mining model based on LightGBM, and in order to improve the classification performance, the model parameters are optimized by grid search algorithm and verified with experiments.

After the selection of important features according to Chapter 2, it is necessary to pre-process the data, according to the visualization analysis, it is found that there is a huge difference between the features, and these data will directly affect the analysis of the results of the final model if they are not processed appropriately. In order to improve the quality of the data, make the data more suitable for mining, display and analysis, so as to avoid affecting the analysis of the results of the multidimensional data mining of teachers' practical teaching cultivation and innovation mechanism, it is imperative to use data preprocessing technology. The next major step is to standardize the dataset after feature selection.

Data standardization is to scale the data proportionally, due to the different features take the value of the range of values there are often differences, which will affect the generalization ability of the model, which will affect the results of data analysis. In order to eliminate the impact of differences in the range of values between features, it is necessary to standardize the features to regulate the inconsistency of the range of data eigenvalues, so this study standardizes the eigenvalues so that they fall into a specific region, which facilitates comprehensive analysis. There are two common methods for data standardization: min-max standardization and z-score standardization.

1) min-max standardization (maximum-minimum method) is a linear transformation of the original data so that the results fall into the [0, 1] interval, which is applicable to the case where the data are distributed in a range.

$$x^* = \frac{x - \min}{\max - \min} \tag{13}$$

where max is the minimum value of the sample and min is the maximum value of the sample.

2) z-score standardization, also known as standard deviation standardization, is to scale the data in accordance with a certain proportion, so that the results fall into a specific interval, after the method of data processing in line with the normal distribution.

$$x^* = \frac{x - \mu}{\sigma} \tag{14}$$

where μ denotes the mean and σ denotes the variance.

The z-score standardization method is suitable for the case where the maximum and minimum values of a feature are unknown or the case of outlier data beyond the range of feature values. According to the principle of z-score standardization method and the use of the scenario, it fits with the data of this study, for which this method is used to standardize the data.

The modeling process of the multidimensional data mining model of teachers' practice teaching cultivation innovation mechanism based on LightGBM algorithm is as follows:

(1) Data preprocessing: data preprocessing is performed on the data generated in the innovation mechanism of teachers' practical teaching cultivation.

(2) Model training: the LightGBM model is initialized, and the preprocessed data are randomly partitioned into a training set and a test set, and the training set is the input to the algorithm in the process of constructing the LightGBM model. The histogram No. method is used to find the optimal segmentation point of the features, and the Leaf-wise leaf growth strategy with depth restriction is applied to generate CART regression trees, the residuals of each regression tree are obtained, and the residuals of the previous round are used as inputs for the next regression tree, and the training is continuous.

(3) Model prediction results: all the training to get the final prediction model, the test set as the input of the model, output the prediction results and evaluate the model.

The model is instantiated according to the modeling process and the default parameters of LightGBM used in the modeling, personalized tuning for different datasets to determine the values, training and prediction of the model, and finally the prediction results.

3. Performance simulation experiments of the model

3.1. Performance evaluation

The selection and use of appropriate evaluation metrics play an important role in understanding the performance of the model, comparing different models, and optimizing the model. In order to ensure the accuracy and practicality of this model, this study chooses the mean square error (MSE), ROC, & AUC for the evaluation and analysis of the model.

The mean square error MSE is mainly used to evaluate the deviation between the observed value and the actual value, y_i is the actual score, n indicates the number of predicted students, the closer the MSE is to 0 indicates the higher the accuracy of prediction and the better the performance of the model.

ROC is an important evaluation metric in model evaluation, which can intuitively reflect the performance of the classifier.The metrics of the ROC curve contain two main ones: true rate (TPR) and false positive rate (FPR). The formula of ROC curve mainly includes the following steps, firstly, a threshold needs to be defined for reflecting the accuracy of the prediction results; the true rate (TPR) and false positive rate (FPR) are calculated, in which TP denotes a true case, FN denotes a false negative case, FP denotes a false positive case, and TN denotes a true negative case; the horizontal axis is the false positive rate, and the vertical axis is the TPR, which is plotted as the ROC curve.

AUC is an important index to reflect the performance of classification model, which is usually used for binary classification problems.The main idea of AUC function is to measure the classification performance of the model by the area under the ROC curve.

$$MSE = \frac{1}{n}\sum_{i=1}^{i}(y_i^{'} - y_i)^2 \tag{15}$$

$$TPR = TP / (TP + FN) \tag{16}$$

$$FPR = FP / (FP + TN) \tag{17}$$

$$AUC = P(p\_\{tpr\} > P\_\{(fpr)\}) \tag{18}$$

In terms of model validation and evaluation, a class in the School of Software at the University of G was used for validation and analysis, the training process of the model is shown in Fig. 1, the mean square error (MSE) of each round was calculated, the closer the MSE is to 0, indicating that the model's performance is better, and when 12 rounds of the process have been trained, the training error in the 2nd round is the smallest, and the validation performance is the best, MSE can only be used as one of the evaluation bases. The ROC graph of the model is shown in Fig. 2, the AUC is about 0.72, and the overall ROC curve is biased towards the y-axis, which indicates that the efficiency of the model is better and has research significance.

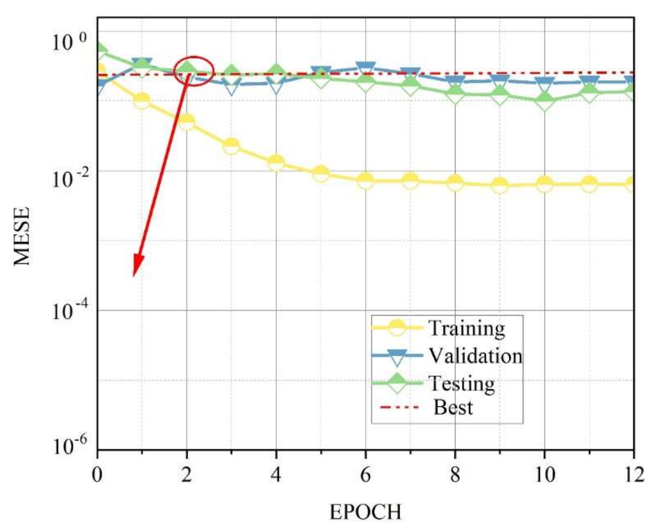

Fig. 1. Training performance

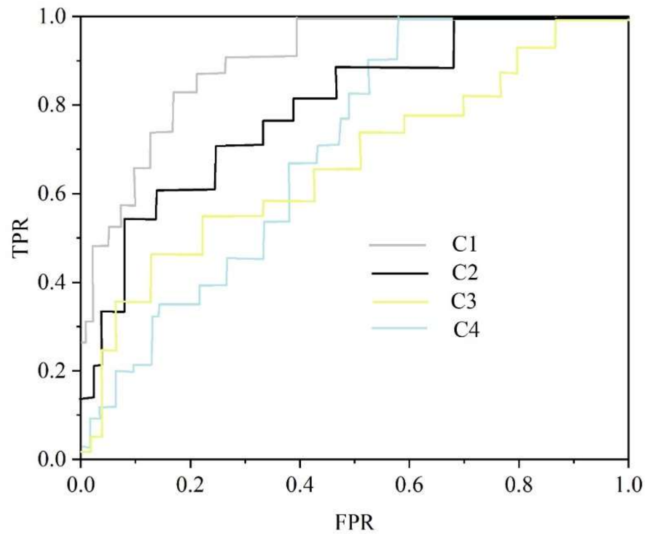

Fig. 2. Roc chart

3.2. Model Error Evaluation

Figure 3 shows the iterative process of this paper's model, it can be seen that when the algorithm is iterated to 40 times, the fitness value reaches about 0.118, which tends to be stable, indicating that this algorithm converges better. This paper chooses the feature selection algorithm based on random forest, the accuracy of the algorithm and the relationship between the number of features is shown in Figure 4, from the figure it can be seen that when the number of features is 5, the accuracy of this paper's algorithm can reach a better level, close to 95%, which illustrates that this paper based on the effectiveness of the feature selection algorithm based on the random forest, and lays the foundation for the construction of the subsequent student behavior and prediction model.

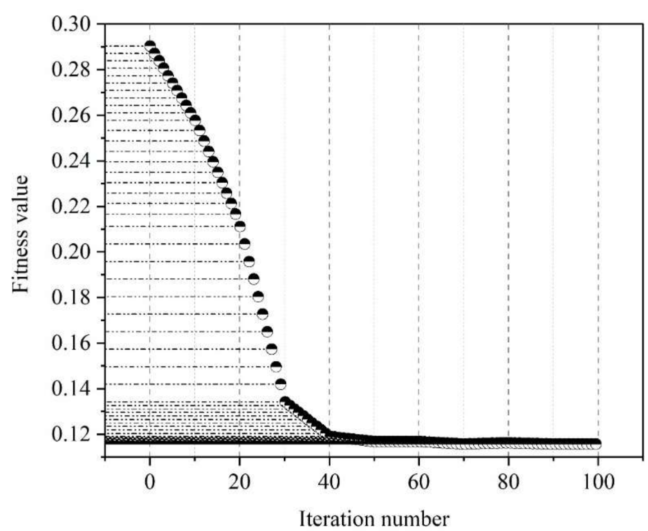

Fig. 3. Algorithm iterative process

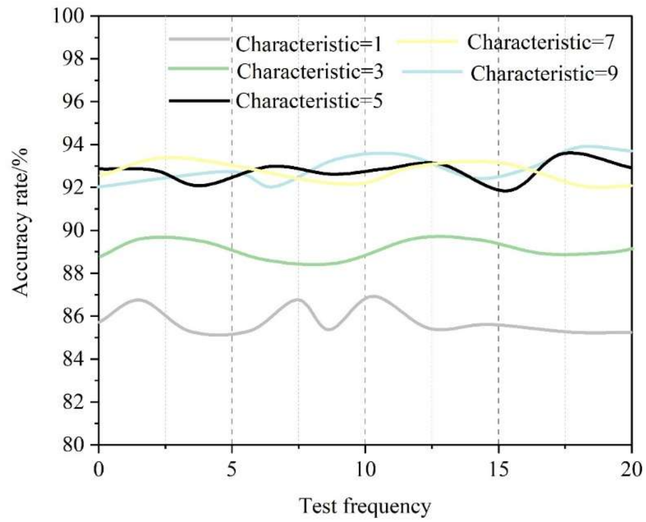

Fig. 4. Feature selection algorithm validation results

3.3. Comparative analysis of models

In order to elucidate the performance of the models implemented under different prediction algorithms in a more detailed and intuitive way, the evaluation indexes were compared at each level of algorithm refinement in the experiments, i.e., the performance of this paper's algorithm and the algorithms such as LSTM, LFFA-LSSVM, and MoE-LSTM were evaluated from four aspects, namely, the accuracy, precision, the average recall, and the F1 value, respectively, and the performances of each algorithm on the same test set is shown in Table 1.

Deepforest model is compared from the accuracy, precision, average recall and F1 evaluation indexes, and the corresponding curve diagram is drawn according to the model performance evaluation data. From the performance comparison, it can be seen that the model in this paper has the optimal performance in the four metrics of accuracy, precision, average recall and F1, and compared with the remaining six algorithms that perform better, LFFA-LSSVM, MoE-LSTM and Deepforest improve the accuracy by 13.07%, 12.17%, 9.73%, and the precision by 12.40%, 10.64%, and 8.16%, optimized 13.10%, 12.24%, and 10.87% in average recall, and improved 12.26%, 10.67%, and 9.67% in F1 value, respectively. Therefore, the model constructed in this paper is better than several other algorithms in terms of structure and performance, which is good feasibility and necessity for the application of the algorithm in the cultivation of innovative mechanisms in practical teaching.

Table 1. Model performance comparison

Algorithm model	Accuracy rate	Precision rate	Average recall rate	F1
LSSVM	0.7717	0.7919	0.7736	0.7827
FA-LSSVM	0.7828	0.8052	0.7784	0.7889
LFFA-LSSVM	0.8256	0.8396	0.825	0.8351
LSTM	0.8037	0.8172	0.8125	0.8155
MoE-LSTM	0.8346	0.8572	0.8357	0.851
Deepforest	0.859	0.882	0.8494	0.861
This model	0.9563	0.9636	0.9581	0.9577

4. Data analysis of innovative management of teachers' practice teaching

There are 49 features in the dataset under the innovative mechanism of teachers' practice teaching cultivation in University of G. Among them, 42 belong to the extracurricular cultivation program outcome credits in the extracurricular practice teaching content of University of G. The other 7 are students' gender, ethnicity, age, class, place of origin, intracurricular grade point and the presence of failing classes. Due to the limitation of the discipline specialization field of the School of Software, many students of the School of Software do not participate in the extracurricular practice program courses, so after excluding the programs in which no student participates, the remaining data containing student credits are the 42 extracurricular cultivation program achievement characteristics in the data set. Among them, there are 5 features in the large category, 15 features in the medium category, and 22 features in the small category.

4.1. Experiments and analysis of different characteristics

Experiment 1 was conducted using a dataset consisting of extracurricular cultivation program outcome broad category features. The dataset features consisted of five extracurricular cultivation program outcome broad category features and seven other basic information for a total of 12 features. Repeating the experiment ten times, the average classification accuracy was 61.8%, and the average importance score of each feature is shown in Figure 5.

In Experiment 1, the scores of scientific research practice activities and grade examination are also higher, while the other three extracurricular practice social practice activities, campus cultural activities and professional development are lower, indicating that conducting scientific research and grade examination is more important to students than other extracurricular activities for their development, so teachers should focus on developing and cultivating students' scientific research and innovation ability to enhance their competitiveness in the grade examination.

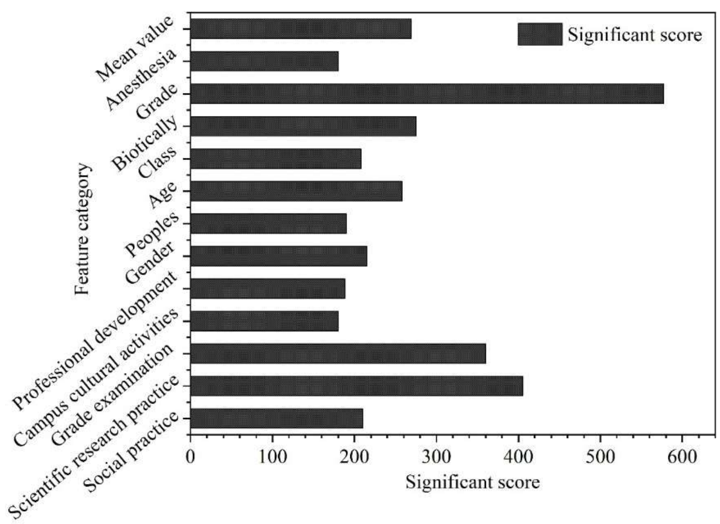

Fig. 5. Experiment one

Experiment 2 was conducted using a dataset consisting of extracurricular cultivation program outcomes in-class features. The dataset features consisted of 15 extracurricular cultivation program outcomes in-class features and 7 other basic information for a total of 22 features. Repeating the experiment ten times, the average classification accuracy is 59.78%, and the average importance score of each feature is shown in Figure 6.

Numbers 1-23 represent vacation social practice, scientific research papers, patent achievements, software achievements, innovative practice, independent scientific research practice, non-professional foreign language proficiency exam, professional foreign language proficiency exam, professional computer grade exam, Chinese language proficiency exam, cultural and sports competitions, cultural and sports activities, academic activities, book report or after-reading, minor in the second degree, gender, ethnicity, age, class, place of origin, in-class GPA, failure or not, average, and other characteristic words and phrases.

Among the mid-category features included in the Research and Practice Activity category, Innovation and Practice has an absolute advantage. Among the medium-class features included in the large category of level examination, non-professional foreign language level examination has absolute advantages. The advantages of these two features need to be analyzed in conjunction with Experiment 3 to analyze their reasons.

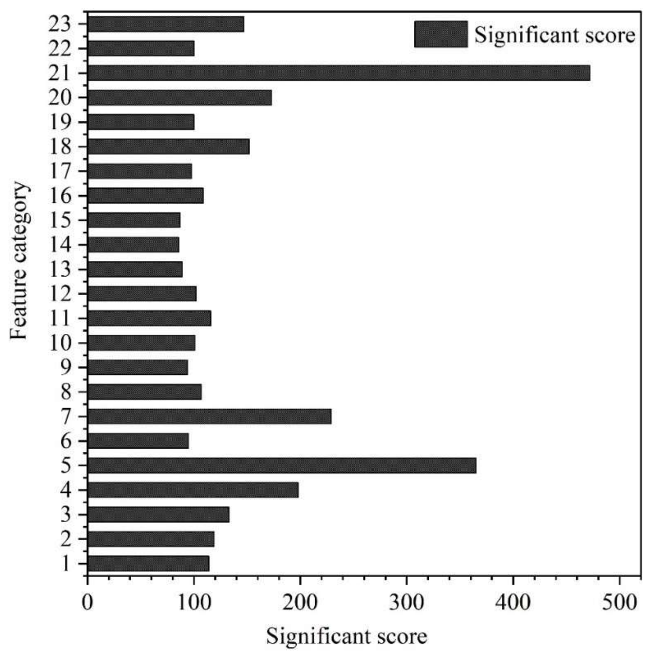

Fig. 6. Experiment two

Experiment 3 Experiment using a dataset consisting of extracurricular incubator program outcome subcategory features. The dataset features consisted of 22 extracurricular cultivation program outcome subcategory features and 7 other basic information for a total of 29 features. Repeating the experiment ten times, the average classification accuracy was 58.96%, and the average importance score of each feature is shown in Figure 7.

The numbers 1-30 in Experiment 3, Experiment 4, and Experiment 5 all represent advanced individuals, academic papers, invention patents, utility model patents, software copyrights, innovation and entrepreneurship training for college students, disciplinary competitions, open innovation experiments, elective experiments, scientific research training, university foreign language level exams, TOEFL, IELTS, GRE, Japanese language professional exams, national computer software proficiency exams, Putonghua exams, sports and literary competition, cultural and sports activities, academic activities, book report or after-reading, minor in second major or second degree, gender, ethnicity, age, class, place of origin, in-class grade point, any failure, average, and other characteristic entries.

Innovation practice contains four subcategories of characteristics, the highest importance is discipline competition, the lowest importance is open innovation experiment, which indicates that the degree of student participation in discipline competition has a great impact on the future development of students, while open innovation experiment has a more limited impact on students who have participated in this activity. Therefore, it can be said that the teaching of open innovation experiment needs to further improve the teaching content to improve the impact on students. Teachers can optimize the cultivation mechanism by starting from students' open innovation experiment teaching link.

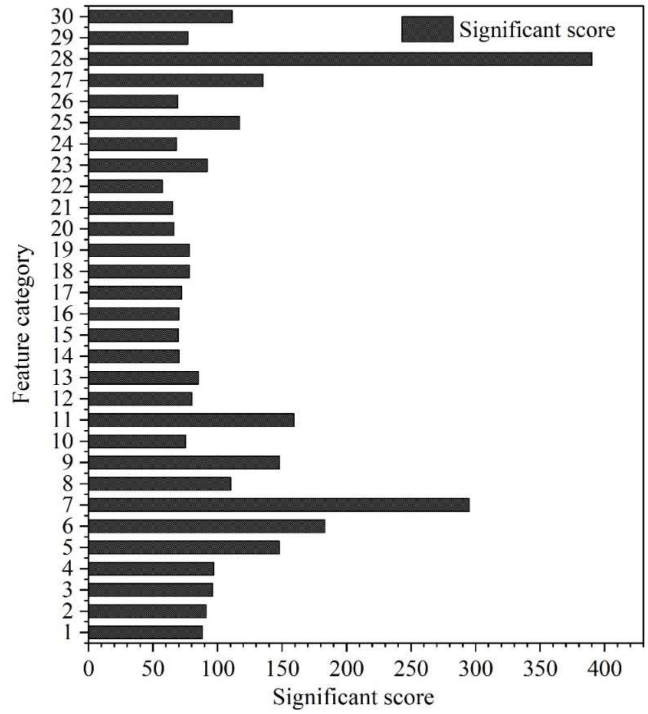

Fig. 7. Experiment three

4.2. Experiments and analysis of different labels

Experiment 4 uses a dichotomous sample set of students with confirmed and undecided destinations at graduation, which consists of 29 features and two types of classification labels, "confirmed destination" and "undecided destination". Repeating the experiment ten times, the average classification accuracy was 69.55%, and the average importance score of each feature is shown in Figure 8.

In Experiment 4, the sample set of "whether to determine the destination at the time of graduation" is studied, and the average value of feature scores is also used as the standard line, and the features located on the standard line are software copyright, college student innovation and entrepreneurship training, disciplinary competitions, open innovation experiments, optional experiments, university foreign language level exams, the origin of the students, and the performance point of the grades of the class. Among them, the place of origin is a condition that students cannot change, but the remaining several high-scoring features indicate that students' efforts to participate in these programs to improve themselves will be very helpful to their future development. Any of the extracurricular formation program outcome characteristics below the standard line that score below 20 can be considered to have little impact on student development, and faculty may appropriately advise students to consider reducing participation in these programs as appropriate.

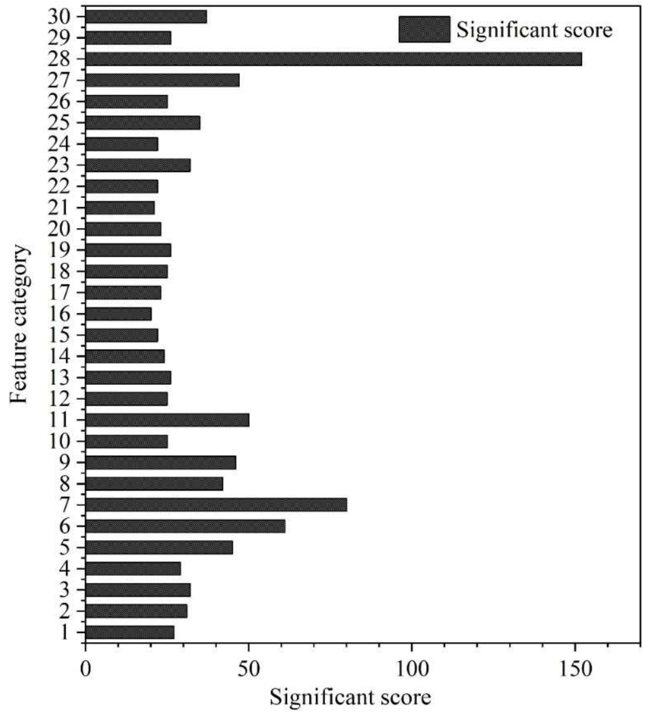

Fig. 8. Experiment four

Experiment 5 uses a dichotomous sample set of students who are pursuing higher education or employment, which consists of 29 features and two types of classification labels, "higher education" and "employment". Repeating the experiment ten times, the average classification accuracy is 80.18%, and the average importance score of each feature is shown in Figure 9.

In Experiment 5, it can be seen that the distribution of feature scores is very different from the previous four experiments. The first rank is no longer the in-class grade point, but the subject competition, which indicates that the subject competition does achieve the role of efficiently selecting different types of students, and it is hypothesized that the students who get high scores in the subject competition are more inclined to choose to go to higher education. The third rank is the university foreign language level examination, which indicates that the English four or six level examination does have a greater impact on the tendency of further education and employment.

For the practical teaching construction of University of G, discipline competitions are very effective in cultivating the ability of students in the School of Software, which highly promotes the development of students' further education. College students' innovation and entrepreneurship training and optional experiments have also made a significant effect on further education and employment. However, although the open innovation experiment has also played a positive role in promoting the development of students, but to a limited extent, teachers need to improve the content of this teaching link, so as to have a stronger impact on the future development of students.

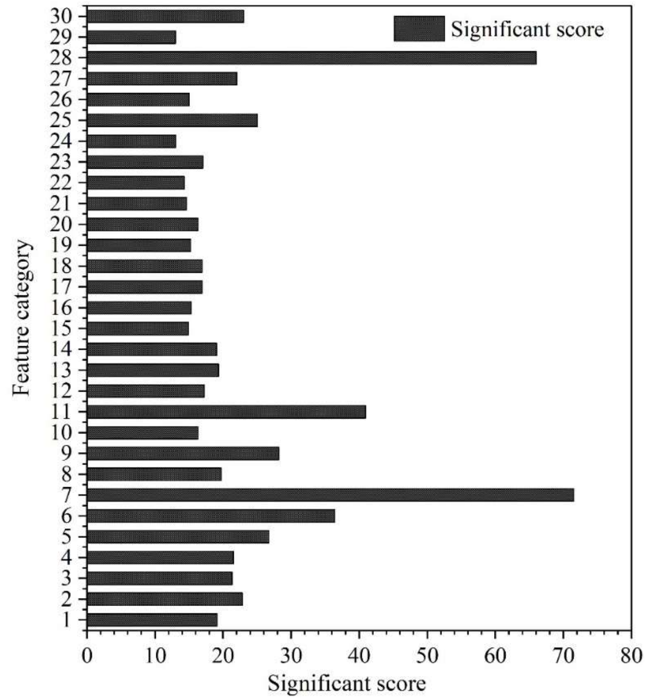

Fig. 9. Experiment five

5. Optimization of Teachers' Practical Teaching Cultivation and Innovation Mechanisms

5.1. Innovative open and innovative experimental course model

Teachers should actively innovate the experimental mode of open innovation experiments to deal with the problem of single experimental mode. First of all, teachers need to have the courage to break the shackles of traditional teaching, jump out of the established experimental framework as well as the process system, and carry out innovative design. The teaching of textbook experiments is essential, but new experimental modes can be adopted, and the content of learning can also be expanded. Secondly, the experimental platform should be innovative. Do not confine students to the laboratory, according to the specific course content, the scene of the experimental course can be moved to the outdoors, so that students can carry out independent exploration and operation. It is also possible to set up a higher degree of freedom to explore the subject, so that students can find information, collect experimental materials, conduct experimental operations and form an experimental report, so as to carry out independent exploration and enhance their innovation ability and disciplinary core literacy.

5.2. Strengthening innovative practice through projects and competitions

The sparrow may be small, but all its organs are complete. Although the student innovation project is small, the school should focus on the project research as a carrier, so that students receive more systematic training, experience the complete process of project research. Through the standardized management of the school, students should not only carry out the principle program design, experimental implementation, data processing and analysis, specimen and prototype production and other specific project research work, but also in the project declaration, opening, mid-term inspection, closing and other aspects of the project to collect information, write the relevant reporting materials, participate in the defense. There are also conducting research, purchasing experimental supplies, reimbursement of funds and so on, each of which is a kind of practical ability training and experience accumulation for students.

5.3. Continuously optimize the practice teaching cultivation and innovation system

Talent cultivation program is the soul of teaching work, the university should always focus on the goal of talent cultivation, to improve the quality of training for the purpose of advancing with the times, in the talent cultivation program constantly highlighting and strengthening the cultivation of innovative ability of students, and the implementation of this into the various aspects of teaching. The university constantly improves the experimental teaching system, and specially supports the experimental teaching reform in the biennial education reform project, focusing on the updating and renovation of the experimental projects in the courses in the undergraduate training program, as well as the addition of innovative experimental projects open to undergraduates, and the transformation of scientific research results and equipment into teaching experiments. The hierarchical experimental teaching system enables students to gradually improve their knowledge and ability through progressive experiments, from theoretical verification and deepening understanding, to comprehensive design and integration of knowledge, and then to innovative practice.

6. Conclusion

In this paper, the feature selection algorithm based on random forest and the data mining model based on LightGBM are designed and applied to the multidimensional data mining of teachers' practice teaching cultivation and innovation mechanism to explore the optimization path of practice teaching cultivation and innovation mechanism.

The feature selection method based on random forest can have a high level of accuracy with a low number of features, which demonstrates the effectiveness of the feature selection method in screening features in this paper, and can significantly improve the prediction accuracy of the data mining model based on LightGBM in the multidimensional data computing task.

In the multidimensional data mining task of fostering innovative mechanisms in teachers' practical teaching, the model in this paper can quickly process multidimensional features, help teachers analyze the problems and phenomena in practical teaching, discover the positive driving effect of open innovation experiments on students' development in higher education, and also discover the limitations of open innovation experiments, prompting teachers to improve the content of that teaching and further promote teaching and learning. Innovation.

The results of data mining in this paper have explored the optimization path of teachers' practical teaching cultivation and innovation mechanism, which can be further optimized from the aspects of innovative curriculum mode, strengthening the practice of competition projects, optimizing the practical teaching cultivation and innovation system, and so on.

Funding

Key scientific research programs of higher education institutions of Henan Province in 2025 (No.: 25B880015) "Research on the Construction and Application of Data-driven OMO Synchronous Teaching Framework: A Case Study of the Teacher Training Course of "Artificial Intelligence Empowering Smart Campus".

References

[1] Hurlbut, A. R. (2018). Online vs. traditional learning in teacher education: a comparison of student progress. American Journal of Distance Education, 32(4), 248-266.

[2] Meng, H., Tang, M., & Wu, M. (2021, November). Current situation on exam-oriented education in China and the outlook for quality-oriented education. In 2021 3rd International Conference on Literature, Art and Human Development (ICLAHD 2021) (pp. 325-331). Atlantis Press.

[3] Korthagen, F. (2017). Inconvenient truths about teacher learning: Towards professional development 3.0. Teachers and teaching, 23(4), 387-405.

[4] Alanoglu, M., Aslan, S., & Karabatak, S. (2022). Do teachers' educational philosophies affect their digital literacy? The mediating effect of resistance to change. Education and Information Technologies, 27(3), 3447-3466.

[5] Gabriel, F., Marrone, R., Van Sebille, Y., Kovanovic, V., & de Laat, M. (2022). Digital education strategies around the world: practices and policies. Irish Educational Studies, 41(1), 85-106.

[6] Taysum, A. (2017). Introduction to the Special Section. External Influences on Education Systems and Education System Leadership. Italian Journal of Sociology of Education, 9(Italian Journal of Sociology of Education 9/2), 1-8.

[7] Lv, K., Yu, A., Gong, S., Wu, M., & Xu, X. (2017). Impacts of educational factors on economic growth in regions of China: a spatial econometric approach. Technological and Economic Development of Economy, 23(6), 827-847.

[8] Li, J., & Xue, E. (2022). Exploring high-quality institutional internationalization for higher education sustainability in China: evidence from stakeholders. Sustainability, 14(13), 7572.

[9] Kaymakamoglu, S. E. (2018). Teachers' Beliefs, Perceived Practice and Actual Classroom Practice in Relation to Traditional (Teacher-Centered) and Constructivist (Learner-Centered) Teaching (Note 1). Journal of Education and Learning, 7(1), 29-37.

[10] Lai, Y. C., & Peng, L. H. (2019). Effective teaching and activities of excellent teachers for the sustainable development of higher design education. Sustainability, 12(1), 28.

[11] Luu, N. Q. H. (2020). Teachers' motivation and its influence on quality education: A study at a center for foreign languages in Vietnam. CTU Journal of Innovation and Sustainable Development, 12(3), 17-26.

[12] Podungge, R., Rahayu, M., Setiawan, M., & Sudiro, A. (2020, June). Teacher competence and student academic achievement. In 23rd Asian Forum of Business Education (AFBE 2019) (pp. 69-74). Atlantis Press.

[13] Gao, R., & Yang, L. (2023). Construction and Practice of a Practical Teaching System for Normal Universities Based on Ability Cultivation. Open Journal of Applied Sciences, 13(10), 1704-1713.

[14] Xu, Y. (2022). Exploration of practical teaching reform based on high-quality international talent cultivation. ICCCM Journal of Social Sciences and Humanities, 1(3), 38-48.

[15] Aglazor, G. (2017). The role of teaching practice in teacher education programmes: designing framework for best practice. Global Journal of Educational Research, 16(2), 101-110.

[16] Simon, E. (2024). Cultivating professional identity: The vital role of practical teaching experience for future educators. Education Sciences, 14(5), 439.

[17] Liu, R. (2018, June). Discussion on the Reform of Practical Teaching and Talent Training Mode in Local Colleges and Universities. In 2018 2nd International Conference on Education, Economics and Management Research (ICEEMR 2018) (pp. 9-12). Atlantis Press.

[18] El Hami Abdelkhalak. (2025). Methods and Applications of Artificial Intelligence: Dynamic Response, Learning, Random Forest, Linear Regression, Interoperability, Additive Manufacturing and Mechatronics.doi:10.1002/9781394351817.

[19] Behnam Amiri Ramsheh,Aydin Larestani,Saeid Atashrouz,Elnaz Nasirzadeh,Meriem Essakhraoui,Ali Abedi... & Abdolhossein Hemmati Sarapardeh. (2025). Toward accurate prediction of carbon dioxide (CO2) compressibility factor using tree-based intelligent schemes (XGBoost and LightGBM) and equations of state. Results in Engineering104035-104035.

[20] Jiasheng Hao,Haomin Xu,Zhinan Peng & Zhen Cao. (2025). An online adaptive ROP prediction model using GBDT and Bayesian Optimization algorithm in drilling. Geoenergy Science and Engineering 213596-213596.