Construction and Analysis of a Decision Tree-Based Predictive Model for Learning Intervention Advice

Chenglong Wang
Department of Education Information Technology
East China Normal University, China
51214108048@stu.ecnu.edu.cn

Abstract

The rapid development of education informatization has accumulated a large amount of data for learning analytics, and adopting educational data mining to find new patterns of data, develop new algorithms and models, and apply known predictive models to the teaching system to improve learning is the challenge and vision of the education field in the era of big data. Learning intervention, as a core concept of learning analytics, refers to the purposeful and planned adoption of direct or indirect strategies or behaviors based on tracking learning behaviors and integrating information about learners' characteristics to give learners personalized guidance and assistance in order to help learners break through the status quo of learning difficulties and improve their learning abilities, so as to achieve tailored teaching. In this study, data mining was conducted on the performance records of students on math problems in an online learning system, and a learning intervention suggestion prediction model was constructed on the basis of decision tree algorithm using Python, with a view to understanding the effectiveness, willingness, style, and other characteristics of the learners' online learning through the analysis results, providing personalized guidance to students, and enabling teachers to intervene with at-risk students and successfully complete the teaching goals. It was found that the most significant impact of the learning intervention advice provided to learners was the number of hints they sought during the learning process, and that learners who needed to be "intervened" or "monitored" could be categorized into two groups: independent inefficient and dependent inefficient according to the model. Therefore, teachers or adaptive learning systems should intervene in a timely and appropriate way for different types of learning crisis groups to solve the problems of poor learning performance, insufficient commitment to learning, poor learning habits, low participation in learning, low self-efficacy and other problems of learners in different learning scenarios.

Keywords: educational data mining, learning analytics, decision tree, learning intervention, predictive model

Introduction

With the in-depth application of information technology in education, the development and change of education informatization has become increasingly significant (Olimov & Mamurova, 2022). At the level of large-scale application, education informatization has experienced changes at the level of Learning Management System (LMS) (Turnbull et al., 2020) and Web 2.0 (Shin & Kim, 2008) application, and the in-depth application of these new technologies has also brought concerns about the explosion of educational data. Students' online learning retains rich information about their learning trajectories, and learning behaviors in social networks show a trend of rapidly increasing data flow, thus a large amount of student learning data is stored in the LMS as the carrier of the environment. These data cannot be captured, stored, managed and processed with typical database software tools within a certain timeframe, and new processing models are needed to have stronger decision-making, insight discovery and process optimization capabilities to adapt to the massive, high growth rate and diverse information assets (Toshniwal et al., 2015). Thus, how to effectively leverage the potential value of big data in education to understand and optimize learning as well as learning contexts is receiving attention and focus from researchers (Daniel, 2019).

Meanwhile, learning analytics uses intelligent data, learner data, and analytical models to discover information and social connections on which to base learning predictions and provide recommendations (Ferguson, 2012). Its essence lies in first discovering the needs of a particular user, using technological methods to acquire data, analyzing the data, helping teachers, students, educational institutions, etc. to interpret the data, and taking interventions based on the results of the data, so as to achieve the goal of improving the effectiveness of learning and teaching (Verbert, 2012). In this context, this study conducted data mining on the performance records of students on math problems in an online learning system, and constructed a learning intervention suggestion prediction model based on the decision tree algorithm using Python, aiming to use the results of the analysis to understand to a certain extent the effectiveness, willingness, style, and other personalized characteristics of the learner's online learning, so that the teacher can carry out personalized interventions and guidance for at-risk students and successfully accomplish teaching goals.

Decision Tree-Based Classification Model for Learning Intervention Advice

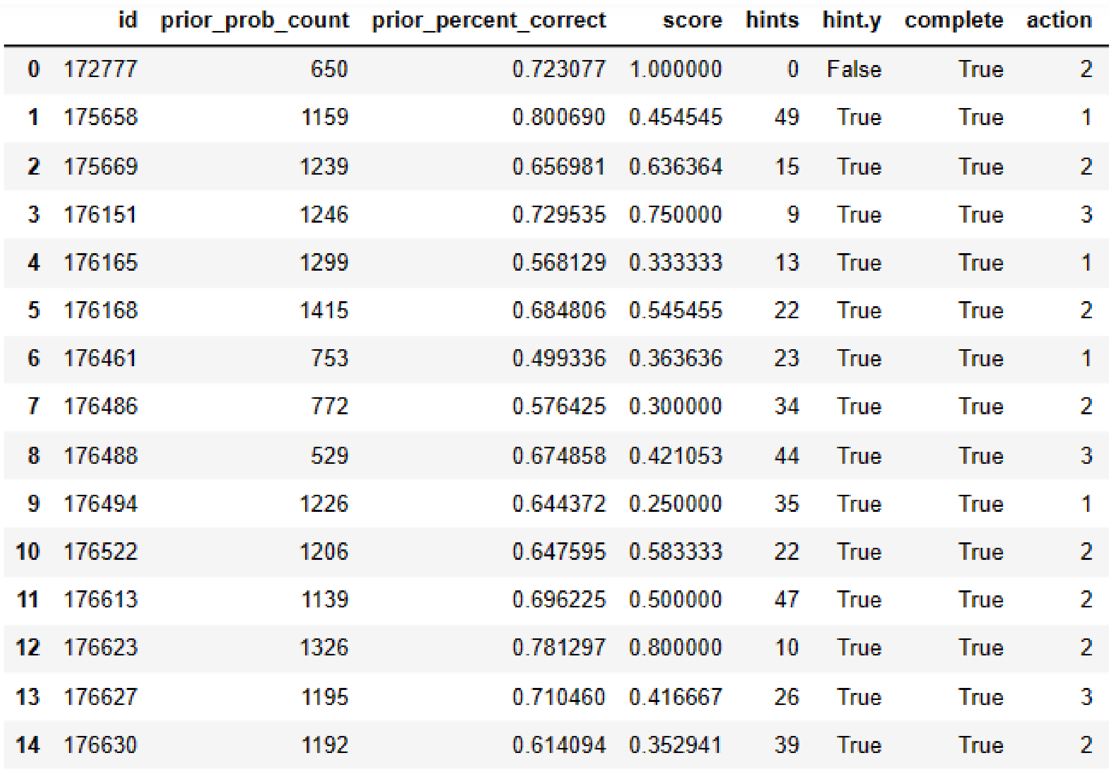
Introduction to Data Classification and Decision Trees:

Classification technique in data mining refers to the categorization of data according to some specified attribute characteristics and has a strong application value (Kesavaraj & Sukumaran, 2013). Classification is the use of training data sets through a certain algorithm to obtain classification rules, the purpose is to obtain a classification function or classification model or called classifier, the model can map the data items in the data set to a given category, based on which the model can be used to identify the category to which the unknown object belongs (Krishnaiah et al., 2014). Classification can thus be used to extract models describing important data classes or to predict future data trends and is the basis of pattern recognition.

Decision tree is a basic but important algorithm in data categorization, it is a tree-like predictive model, usually its internal nodes represent tests on an attribute, while the leaf nodes represent the final category (Maimon & Rokach, 2014). The basic idea of decision tree is to make the original confusing data information gradually clear by dividing the data continuously. Its principles are simple, concrete and close to reality, but are the basis for a series of complex and powerful models (Song & Ying, 2015). Because decision trees are easy to understand and implement, and can produce feasible and effective results for large data sources in a relatively short period of time, this study will utilize the decision tree algorithm to construct a learning intervention recommendation prediction model.

Data description:

The data used in this study came from an online learning system and consisted of records of 378 students' performance on math problems, an overview of the data is shown in Figure 1, and the variables and their meanings are shown in Table 1.



	id	prior_prob_count	prior_percent_correct	score	hints	hint.y	complete	action
0	172777	650	0.723077	1.000000	0	False	True	2
1	175658	1159	0.800690	0.454545	49	True	True	1
2	175669	1239	0.656981	0.636364	15	True	True	2
3	176151	1246	0.729535	0.750000	9	True	True	3
4	176165	1299	0.568129	0.333333	13	True	True	1
5	176168	1415	0.684806	0.545455	22	True	True	2
6	176461	753	0.499336	0.363636	23	True	True	1
7	176486	772	0.576425	0.300000	34	True	True	2
8	176488	529	0.674858	0.421053	44	True	True	3
9	176494	1226	0.644372	0.250000	35	True	True	1
10	176522	1206	0.647595	0.583333	22	True	True	2
11	176613	1139	0.696225	0.500000	47	True	True	2
12	176623	1326	0.781297	0.800000	10	True	True	2
13	176627	1195	0.710460	0.416667	26	True	True	3
14	176630	1192	0.614094	0.352941	39	True	True	2

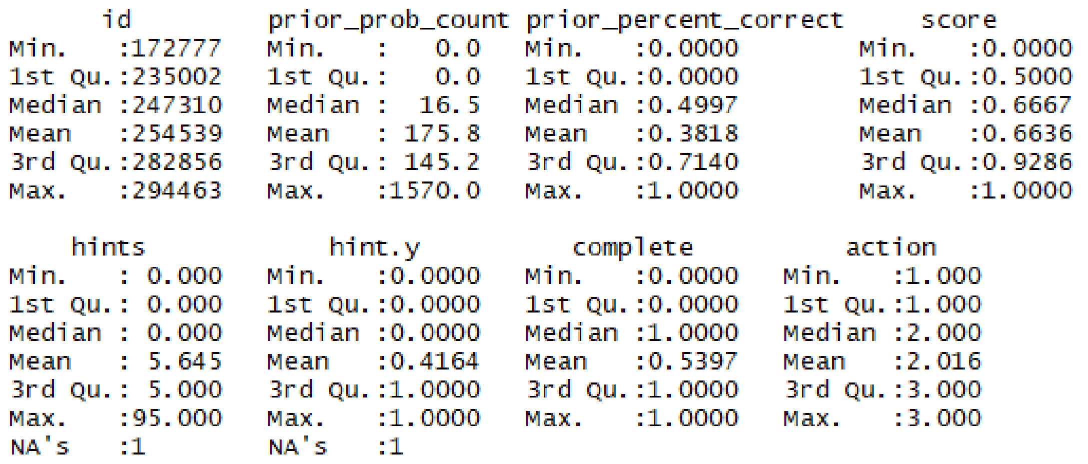
Figure 1. Overview of data

Table 1. Variables and their meanings

Variable name	Meaning
id	Student number
prior_prob_count	Number of questions answered
prior_percent_correct	Percentage of correct answers
score	Grades
hints	Number of hints sought
hint.y	Whether hints have been sought, 1 for yes, 0 for no
complete	Whether a topic has been completed, 1 for yes, 0 for no
action	Types of student behavior, with 1 indicating seeking help from the teacher, 2 indicating

	starting a new topic, and 3 indicating abandonment of the study

Descriptive statistics were done on the data to know the minimum, first quartile, median, mean, third quartile, and maximum values of each variable in all the records and the results are shown in Figure 2.



```
       id          prior_prob_count prior_percent_correct     score       
 Min.   :172777   Min.   :   0.0   Min.   :0.0000        Min.   :0.0000  
 1st Qu.:235002   1st Qu.:   0.0   1st Qu.:0.0000        1st Qu.:0.5000  
 Median :247310   Median :  16.5   Median :0.4997        Median :0.6667  
 Mean   :254539   Mean   : 175.8   Mean   :0.3818        Mean   :0.6636  
 3rd Qu.:282856   3rd Qu.: 145.2   3rd Qu.:0.7140        3rd Qu.:0.9286  
 Max.   :294463   Max.   :1570.0   Max.   :1.0000        Max.   :1.0000  

     hints            hint.y          complete          action     
 Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   :1.000  
 1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000  
 Median : 0.000   Median :0.0000   Median :1.0000   Median :2.000  
 Mean   : 5.645   Mean   :0.4164   Mean   :0.5397   Mean   :2.016  
 3rd Qu.: 5.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:3.000  
 Max.   :95.000   Max.   :1.0000   Max.   :1.0000   Max.   :3.000  
 NA's   :1        NA's   :1                                        
```

Figure 2. Data descriptive statistics

The "score" is the most direct and powerful variable reflecting the online learning effect of learners, so it can be used as the basis for classification, and different learning intervention advice can be provided for learners in different score ranges. The histogram of the frequency distribution of "score" (Figure 3) can visualize the distribution of the number of people in each performance interval, and it is stipulated that learners with a value of "score" less than or equal to 0.3 should be intervened. Learners with a "score" greater than 0.3 but less than or equal to 0.9 should only be monitored, and learners with a "score" greater than 0.9 should not be subject to any intervention (no action).

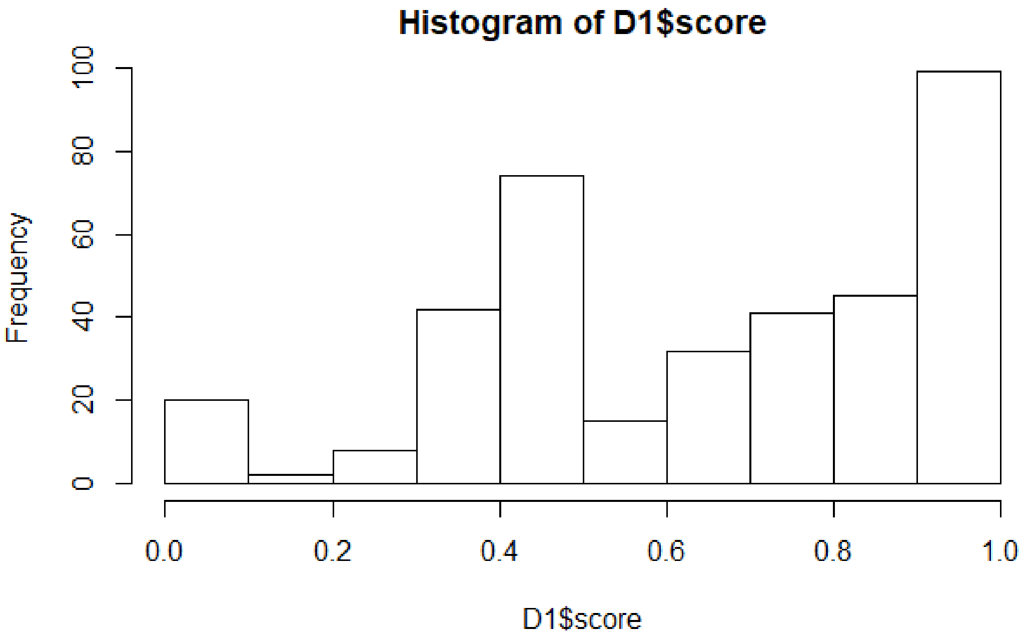


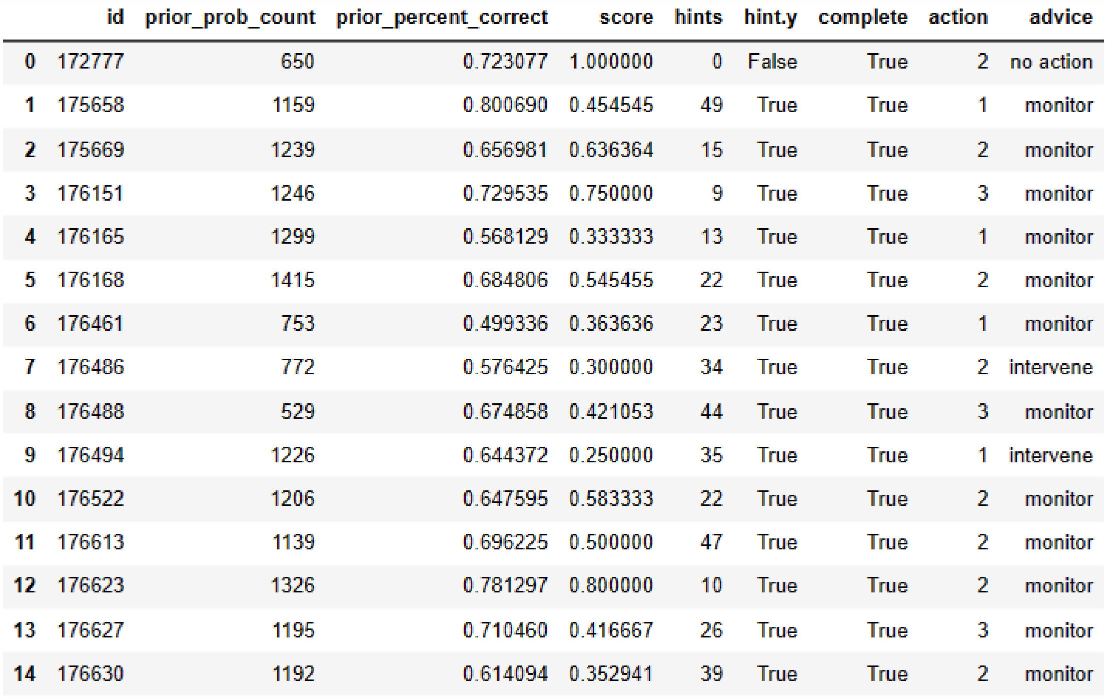
Figure 3. Histogram of the frequency distribution of "score"

Data preprocessing:

In this study, the performance records of 378 students on math problems in an online learning system were used as the training and testing data set. According to the performance intervals combined with the level, form, and purpose of the intervention, the learning interventions that the learners should receive were categorized into "intervened", "monitored", and "no action". The criteria for the classification of learning intervention advice are shown in Table 2, and an overview of the preprocessed data is shown in Figure 4.

Table 2. Criteria for classifying learning intervention advice

Performance interval	Learning intervention advice
[0,0.3]	intervene
(0.3,0.9]	monitor
(0.9,1]	no action



	id	prior_prob_count	prior_percent_correct	score	hints	hint.y	complete	action	advice
0	172777	650	0.723077	1.000000	0	False	True	2	no action
1	175658	1159	0.800690	0.454545	49	True	True	1	monitor
2	175669	1239	0.656981	0.636364	15	True	True	2	monitor
3	176151	1246	0.729535	0.750000	9	True	True	3	monitor
4	176165	1299	0.568129	0.333333	13	True	True	1	monitor
5	176168	1415	0.684806	0.545455	22	True	True	2	monitor
6	176461	753	0.499336	0.363636	23	True	True	1	monitor
7	176486	772	0.576425	0.300000	34	True	True	2	intervene
8	176488	529	0.674858	0.421053	44	True	True	3	monitor
9	176494	1226	0.644372	0.250000	35	True	True	1	intervene
10	176522	1206	0.647595	0.583333	22	True	True	2	monitor
11	176613	1139	0.696225	0.500000	47	True	True	2	monitor
12	176623	1326	0.781297	0.800000	10	True	True	2	monitor
13	176627	1195	0.710460	0.416667	26	True	True	3	monitor
14	176630	1192	0.614094	0.352941	39	True	True	2	monitor

Figure 4. Overview of pre-processed data

Decision Tree-Based Predictive Model for Learning Intervention Advice

Model Building Tools:

Data mining is an effective way to improve the utilization of data and efficiency can be improved by using existing data mining tools such as Eviews, SPSS, SAS, Stata, Matlab, R, WEKA, RapidMiner, etc. all of which have their own distinctive dominant strengths, areas of application, processing capabilities, interface design, security mechanisms, processing efficiency and forms of combination (Romero & Ventura, 2013). This study uses the Python scikit-learn library, which is an open source framework for machine learning and data mining that provides users with a range of simple and effective tools for performing a variety of machine learning tasks such as classification, regression, clustering, dimensionality reduction, model selection, and so on (Pedregosa et al., 2011).

Construction of a Predictive Model for Learning Intervention Advice:

A model or a mapping or a function is learned by means of samples labeled with categories, thus constructing a decision tree capable of predicting recommendations for learning interventions. This process is also known as supervised learning (Niculescu-Mizil & Caruana, 2005) since the labeling of the samples is given artificially. The preprocessed data is subjected to stratified sampling, where 80% is used as the training dataset and the remaining 20% is used as the test dataset. After filtering by feature engineering, the number of questions answered by students "priority_prob_count", the percentage of correct answers "priority_percent_correct", and the number of times seeking hints "hints" as predictor variables and "advice" as outcome variable, the decision tree model obtained is shown in Figure 5.

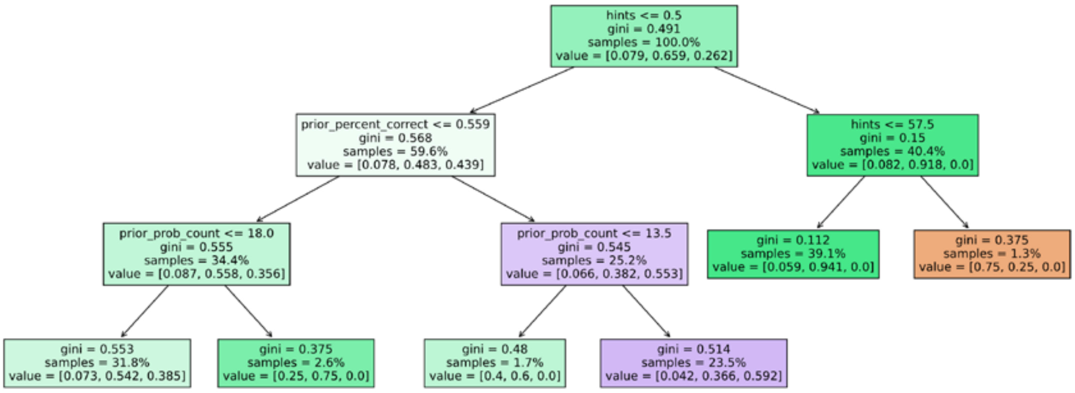


Figure 5. Decision tree-based predictive model for learning intervention advice

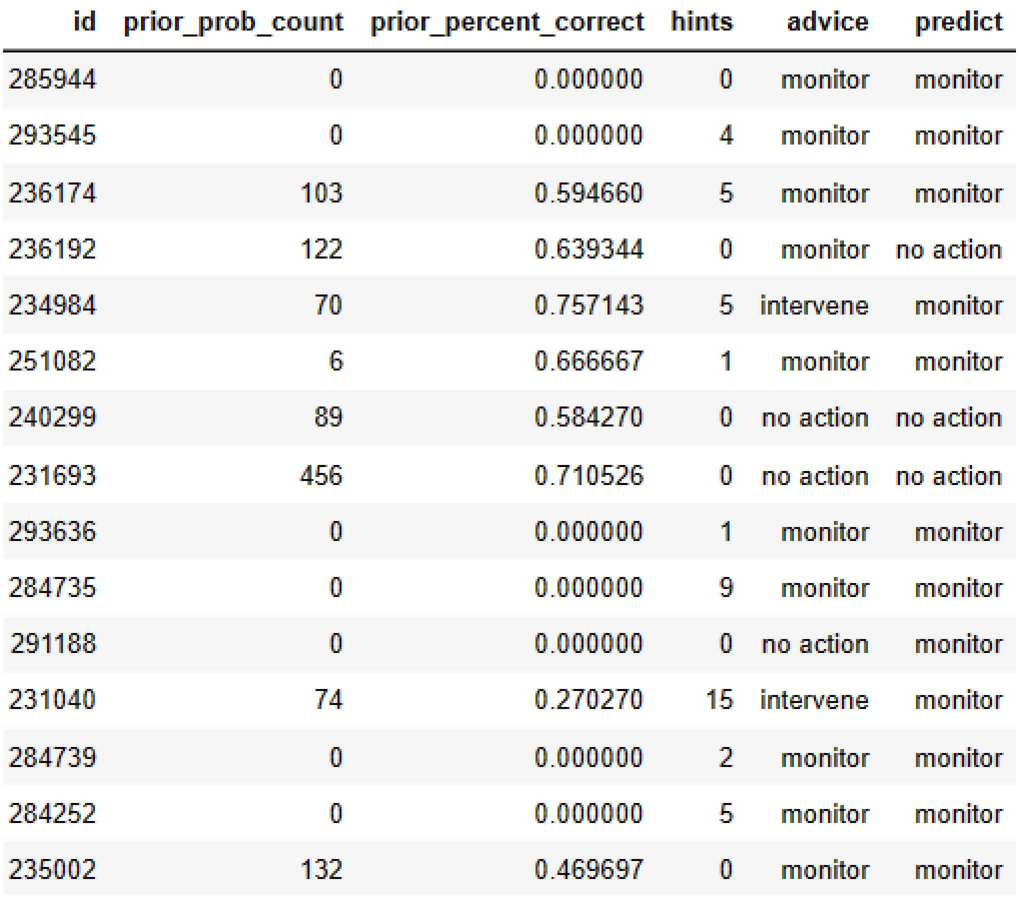
In the figure, the samples on the leaf node represent the proportion of the amount of data used to make judgments and obtain results under this branch in the training dataset. The value in a leaf node indicates, from left to right, the proportion of data in the leaf node that is represented by the three types of learning intervention advice: "intervened", "monitored", and "no action", respectively. In addition, the three feature significance levels, as shown in Table 3, indicate that the number of hints sought by learners during the learning process had the most significant effect on the classified learning intervention advice.

Table 3. Feature significance

Feature name	Significance
hints	0.81
prior_prob_count	0.11
prior_percent_correct	0.08

Model Validation:

The process of arbitrarily classifying an unlabeled sample, given an unlabeled sample, with a model that has been learned, i.e., given its class labeling, is the prediction of the class of an unknown object using a decision tree model. This study generates a prediction of the learning intervention advice corresponding to each record in the test dataset, compares it with the learning intervention advice that the learner should receive according to the grade interval, and obtains a prediction accuracy of 71.05%, and the comparison of the predicted advice with the advice classified according to the grade interval is shown in Fig. 6.



id	prior_prob_count	prior_percent_correct	hints	advice	predict
285944	0	0.000000	0	monitor	monitor
293545	0	0.000000	4	monitor	monitor
236174	103	0.594660	5	monitor	monitor
236192	122	0.639344	0	monitor	no action
234984	70	0.757143	5	intervene	monitor
251082	6	0.666667	1	monitor	monitor
240299	89	0.584270	0	no action	no action
231693	456	0.710526	0	no action	no action
293636	0	0.000000	1	monitor	monitor
284735	0	0.000000	9	monitor	monitor
291188	0	0.000000	0	no action	monitor
231040	74	0.270270	15	intervene	monitor
284739	0	0.000000	2	monitor	monitor
284252	0	0.000000	5	monitor	monitor
235002	132	0.469697	0	monitor	monitor

Figure 6. Predicted advice vs. advice by performance intervals

Result Analysis and Discussion

The construction of a decision tree-based prediction model for learning intervention advice reveals that the most significant influence on the learning intervention advice that should be provided to the learner is the number of hints sought by the learner during the learning process. Whereas learning intervention suggestions are categorized according to performance intervals, the relationship between the number of hints sought and "score" is thus further explored, and the resulting scatterplot is shown in Figure 7. It can be seen that the performance of learners who seek too many hints is less satisfactory, which also confirms that they may have poor abilities in independent learning, adapting to the environment and thinking independently, and it is more necessary for the teacher or the adaptive learning system to analyze the relevant behavioral data of the learner, adopt appropriate intervention strategies, and provide them with targeted support, including activities and resources, so as to ultimately improve the performance of the learner and solve the learning problem (Aleven et al., 2003).

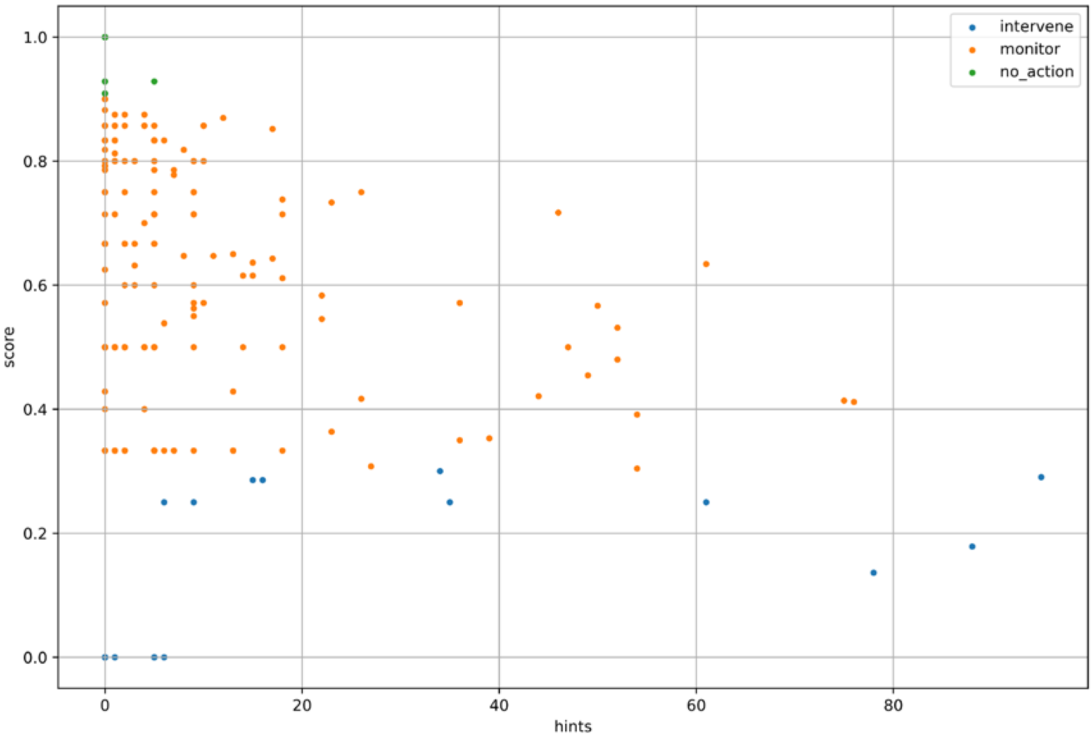


Figure 7. Scatterplot between the number of hints sought and "score"

At the same time, combined with the phenomena reflected in the obtained decision tree model, learners who need to be "intervened" or "monitored" can be categorized into two main groups. The first one is the independent inefficient type, this kind of learners like to learn and think independently, which is manifested in less interaction with the teacher or the elements in the learning environment, but the performance is not satisfactory, for example, the number of times of seeking hints is low but the correct rate of answering the questions is low; the second one is the dependent inefficient type, this kind of learners like to learn in the environment of communicating with the outside world, and they will give feedback to the outside world to get help when they encounter the difficult problems, for example, the number of times of seeking hints is high. Therefore, for the different learning states of learners, teachers can provide manual intervention supported by experience such as psychological counseling, face-to-face meeting, emotional encouragement, etc. (Baugh, 2018), and intelligent agents can provide automatic intervention supported by data such as message reminder, resource recommendation, path planning, etc. (Kabudi et al., 2021), which can bring forward-looking, scientific and differentiated learning support for learners.

Specifically, by looking at the generated decision tree model, it can be observed that different behavioral representations of the learner imply different learning intervention strategies that need to be adopted, which can be broadly categorized as follows:

(1) If learners seek hints very often (leaf node 6), they need to be "intervened", which is generally a precise, personalized one-on-one intervention designed for individuals, and belongs to the high level of intervention. In the online learning environment, if learners frequently seek external prompts and help, indicating that their understanding and cognitive level of knowledge is poor, and that they have difficulty in solving basic problems independently through internalized ways of thinking, then they are at a higher risk of learning than other students, and they are considered to have a more serious learning problem. Therefore, the teacher's intervention for this type of learners tends to be more targeted and directional, through the analysis of the learners' answer situation, to understand the degree of mastery of the learners for each knowledge point, for the weak knowledge points and knowledge of the blind spot recommended learning resources, so that learners to check the leakage of fill in the gaps, to make the learners clearly aware of their own learning status and the existence of the problem, and to provide the learners with a certain learning suggestions and methods.

(2) Learners need to be "monitored " if they seek hints more often (leaf node 5), or if they seek hints less often but answer questions less correctly (leaf nodes 1 and 2), or if they seek hints less often, answer questions more correctly, but answer fewer questions (leaf node 3). This type of intervention is generally a one-to-many intervention for a group, and is a low-level intervention. The "Hawthorne effect" (Sedgwick & Greenwood, 2015) experiment proves that when individuals realize that their behavior is being watched by others, they tend to make positive changes, and the efficiency of learning and interaction increases greatly. That is to say, teachers should participate in the course synchronously, real-time tracking and monitoring of the learners' online learning behavior,

can use the online learning platform message timely feedback class overall answer situation and individual student learning progress bar, coordinating the overall care and individual tutoring, and still give the learners a high degree of autonomy in order to enable them to self-regulate the pace of learning. The main purpose is to prevent learners from having learning crises, to provide warm reminders for learners to enhance their awareness of participating in online learning behaviors, and to help learners rationally arrange their own learning plans (Landrum, 2020).

(3) If the learner seeks fewer hints, has answered more questions, and has a high rate of correct answers (leaf node 4), the strategy of "no action" can be adopted. That is to say, for the cognitive behavioral level with strong self-awareness, initiative and comprehension of the learner, its own or already have a self-driven, self-monitoring, self-feedback and self-regulation metacognitive abilities, tend to learn to learn in the independent learning of the realm of "learning to learn" (Thrun & Pratt, 2012), then there is no need to forcibly interrupt or interfere with the original learning rhythm. Otherwise, it may be counterproductive and not conducive to the cultivation of core literacy and the development of the subjective consciousness of the learner.

Conclusions

In this study, a prediction model for learning intervention advice was constructed on the basis of a decision tree algorithm using Python data mining tools with a dataset of records of students' performance on math problems in an online learning system. It was found that the most obvious influence on the learning intervention advice that should be provided to learners is the number of learners seeking hints during the learning process, and the personalized characteristics such as learning efficiency, question-answering behavior, and learning styles are classified according to the learners' learning efficiency, so as to "prescribe the right medicine", which is of certain positive significance for providing suitable and appropriate learning intervention advice.

Certainly, there are many shortcomings in this study. On the one hand, in the process of data preprocessing, the classification of learners according to performance intervals is subjective, and may not be applicable to other learning interventions in more complex and variable learning situations. On the other hand, the data itself carries less information about the variables, and only some of these variables are selected in the decision tree model construction, perhaps omitting other important feature information, resulting in an incomplete identification of the learner's state (Twala, 2009).

Precisely because intervention is the most direct part of learning analytics technology to improve and enhance learners' learning effectiveness, which is crucial for maintaining learning status, how to carry out timely and appropriate interventions and construct intervention models that can effectively improve learning effectiveness has become an important issue in the field of learning analytics (Kew & Tasir, 2022). Future research hotspots should further delve into the specific learning intervention strategy model based on educational big data, centering on the intervention engine, with the goal of discovering learners' learning difficulties and enhancing learners' learning effectiveness, and starting to build from the four cyclical aspects of learners' learning state identification, intervention strategy matching calculation, intervention strategy implementation, and intervention effect analysis. Based on learning science, teaching theory, curriculum design theory and existing research results, by constructing a multi-dimensional learner portrait, analyzing and monitoring student learning, evaluating the quality of teaching activities, and discovering problems in learning in a timely manner, this is the value of big data and learning analytics in education in the era of information explosion, which should also become a new field of special attention for educational technology researchers (Herodotou et al., 2019).

References

Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help seeking and help design in interactive learning environments. *Review of educational research*, *73*(3), 277-320.

Baugh, A. (2018). The importance of guidance and counseling in present education system: Role of the teacher. *International journal of advanced educational research*, *3*(2), 384-386.

Daniel, B. K. (2019). Big Data and data science: A critical review of issues for educational research. *British Journal of Educational Technology*, *50*(1), 101-113.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5-6), 304-317.

Herodotou, C., Rienties, B., Boroowa, A., Zdrahal, Z., & Hlosta, M. (2019). A large-scale implementation of predictive learning analytics in higher education: The teachers’ role and perspective. *Educational Technology Research and Development*, *67*, 1273-1306.

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, *2*, 100017.

Kesavaraj, G., & Sukumaran, S. (2013, July). A study on classification techniques in data mining. In *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)* (pp. 1-7). IEEE.

Kew, S. N., & Tasir, Z. (2022). Developing a learning analytics intervention in e-learning to enhance students’ learning performance: A case study. *Education and Information Technologies*, *27*(5), 7099-7134.

Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2014). Survey of classification techniques in data mining. *International Journal of Computer Sciences and Engineering*, *2*(9), 65-74.

Landrum, B. (2020). Examining Students' Confidence to Learn Online, Self-Regulation Skills and Perceptions of Satisfaction and Usefulness of Online Classes. *Online Learning*, *24*(3), 128-146.

Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: theory and applications* (Vol. 81). World scientific.

Niculescu-Mizil, A., & Caruana, R. (2005, August). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning* (pp. 625-632).

Olimov, S. S., & Mamurova, D. I. (2022). Information Technology in Education. *Pioneer: Journal of Advanced Research and Scientific Progress*, *1*(1), 17-22.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery*, *3*(1), 12-27.

Sedgwick, P., & Greenwood, N. (2015). Understanding the Hawthorne effect. *Bmj*, *351*.

Shin, D. H., & Kim, W. Y. (2008). Applying the technology acceptance model and flow theory to cyworld user behavior: implication of the web2. 0 user acceptance. *Cyberpsychology & behavior*, *11*(3), 378-382.

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.

Thrun, S., & Pratt, L. (Eds.). (2012). *Learning to learn*. Springer Science & Business Media.

Toshniwal, R., Dastidar, K. G., & Nath, A. (2015). Big data security issues and challenges. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, *2*(2).

Turnbull, D., Chugh, R., & Luck, J. (2020). Learning Management Systems, An Overview. *Encyclopedia of education and information technologies*, 1052-1058.

Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, *23*(5), 373-405.

Verbert, K., Manouselis, N., Drachsler, H., & Duval, E. (2012). Dataset-driven research to support learning and knowledge analytics. *Journal of Educational Technology & Society*, *15*(3), 133-148.