ISSN: 2664-9640

Student Achievement and Cognitive Behavior Prediction Using Data Mining Techniques

Huan Jie

University of Macau, Macau, China.

Abstract

When studying students' reading achievement and literacy, data mining technology is becoming a popular research tool. Most studies of student performance prediction are based on the analysis of students' behavior as well as the thesis about learning from students' emotional attitude, cognitive learning and history analysis is less. Furthermore, the characteristic of a single data dimension in student performance prediction research constraints the possibility of analyzing student literacy from multiple dimensions. Therefore, this paper studies students' reading literacy from three dimensions of historical learning behavior characteristic, affective cognitive characteristic of learning reading and learning behavior characteristic. In this paper, four prediction and classification techniques — decision tree model, random forest, support vector machine, artificial neural network and discriminant analysis — are used to build prediction and classification models, and the effects of modeling are especially in comparison and scrutinized. The data set is derived from the 2018 PISA test data of Chinese mainland students from the Organization for Economic Cooperation and Development (OECD). After the models are established, they are evaluated by confusion matrix, gain curve, ROC curve, AUC value and other model evaluation indicators. The classification efficiency of all models is above 85%, and the accuracy of C5.0 algorithm is up to 92%. Following the evaluation of the model's effect, this paper continued to rank the importance of classification variables, analyzed the top three most important indicators, and recorded the indicators with little impact. The results show that students' own cognition of learning reading is of great significance to learning reading, and history learning behavior is also a key factor affecting students' reading performance, while e-book reading has little impact on reading literacy. The research results of this paper show that data mining model can be applied to PISA questionnaire database, which provides ideas for further research.

Keywords

Model Evaluation，Decision Trees，Reading Literacy，PISA，Emotional and Cognitive Characteristic.

1. Introduction

Nowadays, everyone lives in an information-overload world, and the ways in which information is presented in written form are expanding as well. People are expecting that the information can be used more quickly. Therefore, reading ability is especially important. Reading and literacy are not only the foundation for success in scientific research and related areas, but they are also required for success in all aspects of life. In this context, it is particularly important to evaluate students' reading literacy, and the evaluations are of profound significance.

In the international student ability evaluation plan, the Organization for Economic Cooperation and Development (OECD) regards reading as the primary test and evaluation purpose, and it has provided useful data on reading literacy assessments for students around the world at the

ISSN: 2664-9640

age of 15, when compulsory education ends. The assessment's category includes text information extraction, interpretation, and integration, as well as the ability of the summary. How to use the large amounts of PISA survey data to improve prediction and analysis of the student's reading ability has become one of the major challenges for education workers. To assess student achievement in education and the level of learning, the primary methods are graph display and descriptive statistics, but these basic methods lack in-depth excavation of the hidden information capacity[1].Simultaneously, data mining technology is becoming more widely used in student performance and ability evaluation, owing to its effectiveness in education management, prediction of student performance, and analysis of education influencing factors. As a result, education data mining has emerged as a new research field. Education data mining (EDM) examines the data of a variety of education stakeholders, such as students, teachers, education administrators, and educational institutions, using original education data to draw meaningful insights.

In 2007 Emmanul N. Ogor classified students by their scores in five subjects[2]， In 2017, Suhas S Athani used Bayesian network and other models to predict students' 2015 PISA mathematics test scores and students' social performance[3]， Farid Jauhari et al. classified students' academic performance by establishing decision trees in 2019[4]. Data mining technology has been widely used in student performance prediction and classification, which is mainly based on academic achievement forecasts. As a result, there are few studies on predicting students' achievement based on their inner psychological state. This paper will add the psychological state and cognitive factors based on students' historical scores, then classify and predict students' scores, and compare and evaluate the prediction models. In terms of the selection of prediction methods, this paper chooses several most commonly used machine learning models for classification and prediction, including decision tree, random forest, neural network, support vector machine and discriminant analysis, which provides basic ideas for data mining methods of PISA questionnaires in the future.

This paper's subsequent content includes the four following elements: It begins by introducing the relevant classification model, then explains the data processing process and the data and variables to be used, after that evaluates and analyzes the model operation results, and finally summarizes the analysis conclusions and presents the prospects.

2. Classifications Models

Classification is one of the most common applications of data mining. The main purpose of classification is to obtain the relationship between different attributes of data through training sets, to summarize the method to distinguish data categories establish a classification rule with variables through machine learning and predict the new data categories. Among the classification algorithms of data mining, the most common classification models are decision tree, neural network, support vector machine and Bayesian network, among others. In this study, several decision tree algorithms, neural network, discriminant analysis and support vector machine are selected as the most convenient classification algorithms. Here is a brief introduction [11].

C5.0 algorithm is born from information theory. Information transmission is realized through a transmission system composed of information source, information channel and home. The information transmission process is sent by information source, transmitted by information channel and received by home. If the sent information is denoted as U and the received information is denoted as I, the model is $P(U|V)$. It also known as the channel transmission probability matrix, denoted as:

ISSN: 2664-9640

$$\begin{bmatrix} P(u_1|v_1) & P(u_2|v_1) & \cdots & P(u_r|v_1) \\ P(u_1|v_2) & P(u_2|v_2) & \cdots & P(u_r|v_2) \\ \vdots & \vdots & & \vdots \\ P(u_1|v_q) & P(u_2|v_q) & \cdots & P(u_r|v_q) \end{bmatrix}$$

*I*n the above formula, $P(u_i|v_i)$ *r*epresents the probability that the host receives the message v_i *a*nd the message source u_i sends the message. The amount of information can be expressed as:

$$I(u_i) = \log_2 \frac{1}{P(u_i)} = -\log_2 P(u_i)$$

The amount of information is in bits so it's logarithmic base 2.

Information entropy is the mathematical expectation of the amount of information. It represents the average uncertainty before the information is sent by the source, and its mathematical definition is as follows:

$$Ent(U) = -\sum_i P(u_i) \log_2 P(u_i)$$

When the probability distribution of signal *U i*s known and the signal $V = v_i$ is received, the probability distribution of the sent signal changes, so the average uncertainty of the source becomes:

$$Ent(U|v_j) = -\sum_i P(u_i) \sum_j P(u_i|v_j)$$

It is called posterior entropy. Since the received signal *V i*s a random variable, the expectation of posterior entropy is:

$$Ent(U|V) = \sum_j P(v_i) \sum_i P(u_i|v_i) \log_2 \frac{1}{P(u_i|v_i)} = \sum_j P(v_j)(-\sum_i P(u_i|v_j) \log_2 P(u_i|v_j))$$

So, $Gains(U,V) = Ent(U) - Ent(U|V)$ *i*s called information gain.

The information gain rate is based on the information entropy value divided by the information entropy of independent variable *V.* The information gain rate avoids the influence of the number of categories in a signal group on the information gain, and its calculation formula is as follows:

$$GainsR(U,V) = Gains(U,V) / Ent(V)$$

The decision tree regards the target variable as the information *V s*ent by the information source, and the input variable as the information received by the home. The target variable is random to the home, and the information entropy is used to represent its average uncertainty. Finally, the variable with the maximum information gain rate is selected as the best grouping variable. After we find the best grouping variable and the best segmentation point of the sample set through training, and finally generate a decision tree. According to the constructed decision tree, the classification rules are extracted, and the new data sets are classified.

Different from C5.0, the target variables of CART decision tree model can be discrete or continuous. Gini coefficient and variance are used as the basis for selecting variables in CART, and Surrogate is used to process real variables.

The CHAID model is similar to the algorithms of the first two models, but the uniqueness of CHAID is that it determines the current optimal grouping variables and segmentation points from the perspective of statistical significance test.

ISSN: 2664-9640

Random forest is an algorithm that integrates multiple decision trees through the idea of Ensemble Learning. Its basic unit is decision tree, and its essence is an Ensemble Learning method. Each decision tree is a classifier. Randomly selected samples generate different decision trees and finally get different classification results. In this process, random forest realizes sample randomness and feature randomness. The random forest then aggregates the results of all the categories, designating the category with the most votes as the final output, which is the simplest Bagging idea.

Artificial neural network (ANN) is a modeling method that uses computer to simulate human brain thinking. Biological neurons are simulated as processing units, and organic connections of processing units are used to solve classification and prediction problems. The bottom layer of neural network is called the input layer, the middle layer is the hidden layer, and the upper layer is the output layer. The number of layers and the number of units in each layer determines the complexity of neural network. The establishment of neural network goes through the process of data preparation, network structure determination and network weight determination. The advantage of neural network algorithm is that it can process noisy data and can be used for modeling any complex pattern, so it is suitable for classification problems.

Support vector machine (SVM) is the result of statistical methods, is divided into support vector regression and support vector classification machine, the former type is used to predict the input variables and the numerical model of the target variable relations and forecast, which is used to study the type of input variables with binary target variable and predict the relationship between, to deal with the problem of classification of input variables and has many advantages. LSVM input linearly separable training set and output separate hyperplane and classification decision function. This method can obtain the maximum interval hyperplane, so the unique solution can be obtained.

Discriminant analysis is also a classical multivariate statistical analysis method, which can determine the relationship between input variables and predicted variables through existing data. Similarly, discriminant analysis focuses on the prediction of categories, so it can be used for classification and prediction.

Above is a brief introduction of several machine learning classification algorithms.

3. Data Collection And Variables

The introduction of data set structure is divided into two parts, one is vertical introduction, one is horizontal introduction. First, from the perspective of horizontal introduction, in this paper, we selected 11,704 students from Mainland China who took the test in 2018[5]. They came from Beijing, Shanghai, Jiangsu and Zhejiang provinces. From the vertical perspective, Pisa questionnaire setting [6] sets questions including family background, language learned at school, views on reading, learning time and learning plan and so on. In 2018's PISA questionnaire, it involves three modules in the questionnaire design framework of literacy assessment, namely Student background, Schooling Constructs and Non-cognitive/Metacognitive Constructs. There are evaluation indexes about out of school reading experience in Student background. Schooling Constructs are evaluated in two aspects: Teacher qualifications and professional development. Non- cognitive/Metacognitive Constructs is described by indicators in the reading related outcomes; altitudes; motivation and strategies dimension [7]. Because this paper is studied by students as the starting point, Student background and Non-cognitive/Metacognitive Constructs are extracted from all the questionnaire frameworks for the selection of indicators. In this paper, data variables are constructed from three dimensions to achieve the purpose of research.

ISSN: 2664-9640

Further, this paper studies the classification and prediction of reading scores. Therefore, eight indicators including three dimensions are selected as influencing factors of students' reading literacy. In order to explore the relationship between students' learning characteristics and reading level from the perspective of cognition, the index system is constructed into three dimensions: History learning behavior characteristics, learning cognitive characteristics, learning behavior characteristics, The aim variable is the last row: whether the student's performance can be better than the global average of student reading literacy assessment, which is 487. Then, this paper uses decision tree, neural network, support vector machine and other data mining methods to study the influencing factors of students' reading literacy, classify and predict, and select the optimal model.

The data studied in this paper take the questions of the questionnaire as the variables of the study, and the options of the questionnaire as the values contained in each variable. There is one continuous variable and seven discrete variables in the data. The variable values of Q1, Q2, Q4 and Q5 represent quantity, the variable values of Q6, Q7 and Q8 represent degree, and Q3 is a discrete variable. Finally, whether the score can be higher than the average value of the world students' reading literacy assessment is a discrete variable, divided into "yes" and "no"[10].

The eight variables included are explained in detail in the following table:

Table1: variables details

Question number	Question Description	Probable values	Interpretation of variables
Student historical performance characteristics			
Q1(ST127Q03TA)	Have you ever repeated a <grade>? At <ISCED 3>	{1,2,3}	1= no never 2= yes once 3=yes, twice or more
Q2(ST062Q03TA)	In the last two full weeks of school, how often: I arrived late for school.	{1,2,3,4}	1=never 2=one or two times 3=three or four times 4=five or more times
learning behavior characteristics			
Q3(LMINS)	Learning time (minutes per week) - <test language>		Continuous variable
Q4(ST150Q01IA)	During the last month, how often did you have to read for school: Texts that include diagrams or maps	{1,2,3,4}	1=many times 2=two or three times 3=once 4=not at all
Q5(ST012Q08NA)	How many in your home: E-book readers (e.g. <Kindle>, <Kobo>, <Boo keen>)	{1,2,3,4}	1=none 2=one 3=two 4=three or more
Student cognitive characteristics			

ISSN: 2664-9640

Q6(ST165Q03IA)	Usefulness for writing a summary: Before writing the summary, I read the text as many times as possible.	{1,2,3,4}	1=none 2=one 3=two 4=three or more
Q7(ST164Q05IA)	Usefulness for understanding and memorizing text: I summaries the text in my own words.	{1,2,3,4,5,6}	1=not useful at all 2= (2) 3= (3) 4= (4) 5= (5) 6=very useful
Q8(ST208Q04HA)	How true for you: My goal is to understand the content of my classes as thoroughly as possible.	{1,2,3,4,5}	1=not at all true of me 2=slightly true of me 3=moderately of me 4=very true of me 5=extremely true of me
Student scores (Target)	Whether student performance is above international standards	{yes, no}	"yes" =Higher than the international average "no" =Lower than the international average

4. Results And Discussions

In the research, there are seven classification models built[12]: three decision tree algorithms, one random forest, one artificial neural network, one support vector machine, and one using discriminant analysis. In this paper, the accuracy of the test data of these models is discussed.

4.1. Output the comprehensive evaluation result of the model, which is the correct rate of model training and testing.

The results show that the correct rate of the C5.0 algorithm is the highest whether it is the training set or the test set[13].

The full name of AUC is Area Under the Curve of ROC, which is the area under the ROC curve. AUC is defined as the area under the ROC curve. The AUC value is often used as the evaluation criterion of the model because in many cases the ROC curve does not clearly indicate which classifier performs better, and as a value, a classifier with a larger AUC performs better[9].

Table2: Model accuracy

	Model accuracy	AUC
CHAID	85.99%	0.764
C5.0	86.21%	0.622
CART	85.99%	0.576
ANN	85.93%	0.729
SVM	85.81%	0.671
DA	68.67%	0.705

ISSN: 2664-9640

Among the prediction results of all the classifiers, C5.0 algorithm has the best classification prediction effect, and DA algorithm has the worst classification effect. In terms of comparison of AUC values, CHAID algorithm has the best classification effect. The classification effect of decision tree on PISA questionnaire is higher than other classification methods.

4.2. Model Parameters and Confusion Matrix Results:

After establishing the model with machine learning algorithm and statistical methods, it is necessary to evaluate the quality of the model. The accuracy rate, accuracy rate, recall rate, specificity and other indicators can be used to evaluate the quality of the classification results. Since the purpose of this study is a dichotomous problem, it can be assumed that the values of class variables are positive (P) and negative (F). If the classification result is correct, it is recorded as positive and negative. Is properly labeled as positive real positive (P) is true positive (TP), falsely marked positive for the actual negative (N) is true positive (TP), on the contrary, be negative for the actual error sign positive (P) for false negatives (FN), is properly labeled as negative real positive (P) for false positives (FP). After obtaining these indicators, accuracy rate, accuracy rate, recall rate and specificity can be calculated. Sensitivity can indicate whether the model can accurately identify the predictive variables[10].

Table3: Confusion Matrix

Predict Class				
Actual Class		No	Yes	Total
	No	TN	FP	N
	Yes	FN	TP	P
	Total	N'	P'	

Precision is the proportion of correct model prediction under the condition that the model result is positive, accuracy is the proportion of the classification model with more correct judgment results in the total results, specificity is the proportion of model prediction when the model truth value is negative. These indicators will be presented in the following mixed matrix and calculated by the following formula:

In the construction process of CART model, the minimum impurity change is set to 0.0001, the impurity measurement of classification target is set to Gini, and the over-fitting prevention set (%) is set to 30%. In the Confusion matrix results of CART model, there are 95 TN, 4887 TP, 746 FP and 66 FN as the classification result. The confusion matrix results of CART model are shown in the following table:

Table4: CART Confusion matrix

CART Confusion matrix			
	no	yes	Total
no	95	746	841
yes	66	4,887	4953
Total	161	5633	5794

C5.0 model uses partitioned data and builds a model for each partition without building branches. The model is built using the original algorithm. In the Confusion Matrix results of C5.0 model, there are 91 TN, 4953 TP, 750 FP and 49 FN as the classification result. The running results are as follows:

ISSN: 2664-9640

Table5: C5.0 Confusion matrix

C5.0 Confusion matrix			
	no	yes	Total
no	91	750	841
yes	49	4,904	4953
Total	140	5654	5794

During the construction of CHAID model, the significance level of segmentation was set to 0.05, the significance level of merger was set to 0.05, and the significance value was adjusted by Bonferroni method. The chi-square used for the category target was set as Pearson, and the maximum number of iterations of convergence was 100. In the Confusion Matrix results of CHAID model, there are 95 TN, 4887 TP, 746 FP and 66 FN as the classification result. The mixed matrix of model running is shown in the following table:

Table6: CHAID Confusion matrix

CHAID Confusion matrix			
	no	yes	Total
no	95	746	841
yes	66	4,887	4953
Total	161	5633	5794

The number of Random trees constructed by the Random Tree model is 100, the sample size is 1.0, the maximum node number is set to 10000, the maximum Tree depth is set to 10, and the size of the smallest node is set to 5. Moreover, the Random Tree model will not be constructed when the accuracy can no longer be improved. In the Confusion matrix results of Random Tree model, there are 421 TN, 3708 TP, 1165 FP and 402 FN as the classification result.

Table7: Random Tree Confusion matrix

Random Tree Confusion matrix			
	no	yes	Total
no	421	1165	1586
yes	402	3708	4110
Total	823	4873	5696

The regression accuracy of LSVM model is set to 0.1, and the sorting order of classification targets is ascending. In order to use regularization measures to limit the ability of the model and avoid the problem of non-differentiability when the absolute value function is evaluated again, the absolute value sum is changed to the sum of squares, so the L2 norm penalty is selected, and the penalty parameter is se*t* to 0.1*[10].* In the Confusion matrix results of LSVM model, there are 14 TN, 4949 TP, 827 FP and 4 FN as the classification result.

Table8: LSVM Confusion matrix

LSVM Confusion matrix			
	no	yes	Total
no	14	827	841

ISSN: 2664-9640

yes	4	4,949	4953
Total	28	5773	5794

ANN algorithm constructs a three-layer neural network with a hidden layer, including six hidden layers. In the Confusion matrix results of ANN model, there are 92 TN, 4887 TP, 749 FP and 66 FN as the classification result.

Table9: ANN Confusion matrix

ANN Confusion matrix			
	no	yes	Total
no	92	749	841
yes	66	4,887	4953
Total	158	5636	5794

*I*n the Confusion matrix results of DA model, there are 519 TN, 3460 TP, 322 FP and 1493 FN as the classification result.

Table10: DA Confusion matrix

DA Confusion matrix			
	no	yes	Total
no	519	322	841
yes	1493	3460	4953
Total	2012	3782	5794

Table11: Performances of classifier models

	Recall	Precision	Accuracy	Specificity
CHAID	0.8870	0.8676	0.8666	0.9867
C5.0	0.8918	0.8674	0.8688	0.9901
CART	0.8870	0.8676	0.8666	0.9867
ANN	0.8906	0.8671	0.8661	0.9867
SVM	0.9834	0.8568	0.8633	0.9992
DA	0.3829	0.9149	0.6921	0.6986
Random Tree	0.7346	0.7609	0.7126	0.9022

4.3. Predicted results of variable importance

There are four important variables in the output of the results calculated by the CART model. The first two important classification variables are "Usefulness for understanding and memorizing text and Usefulness for writing A Summary ", and the third grouping variable is " learning time ".

The C5.0 model outputs the four most important variables. The first one that affects students' reading literacy is "Using for understanding the text: Memorizing "is the category of reading cognition. The second important influencing factor is" for writing a summary ", which is also the cognition of reading methods. The third is learning time.

ISSN: 2664-9640

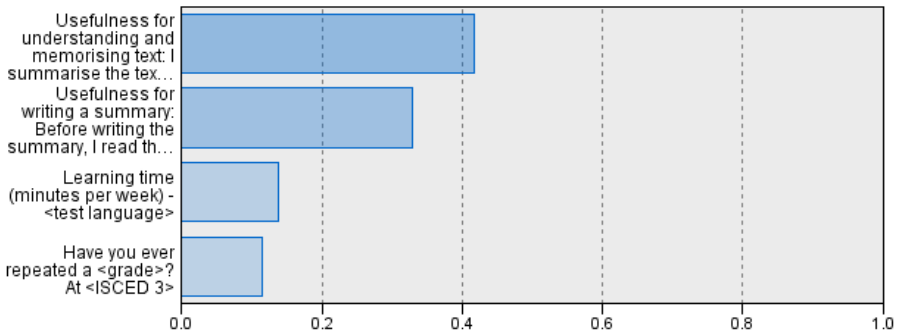


Figure2: C5.0 model results

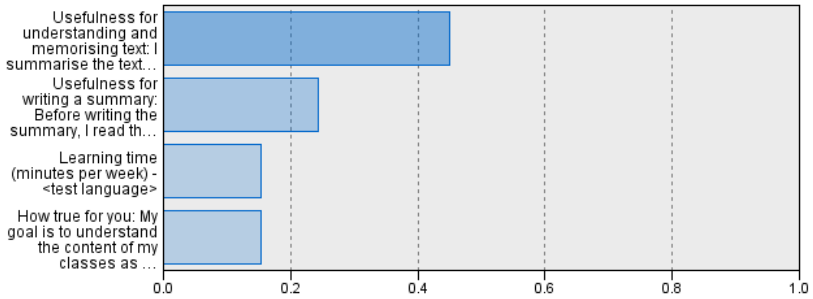


Figure1: CART model results

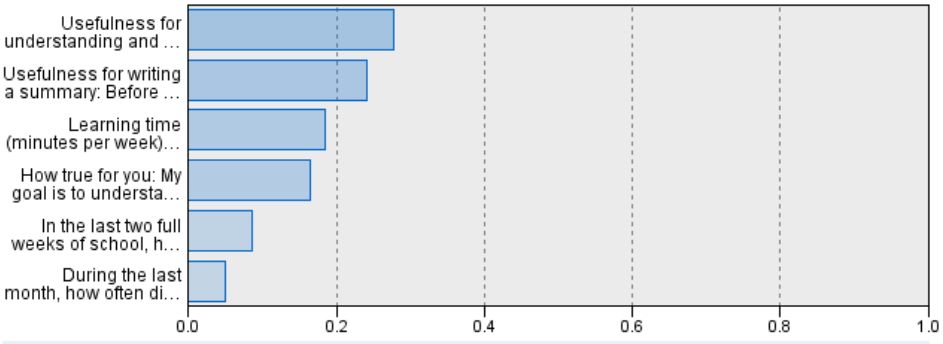


Figure3: CHAID model results

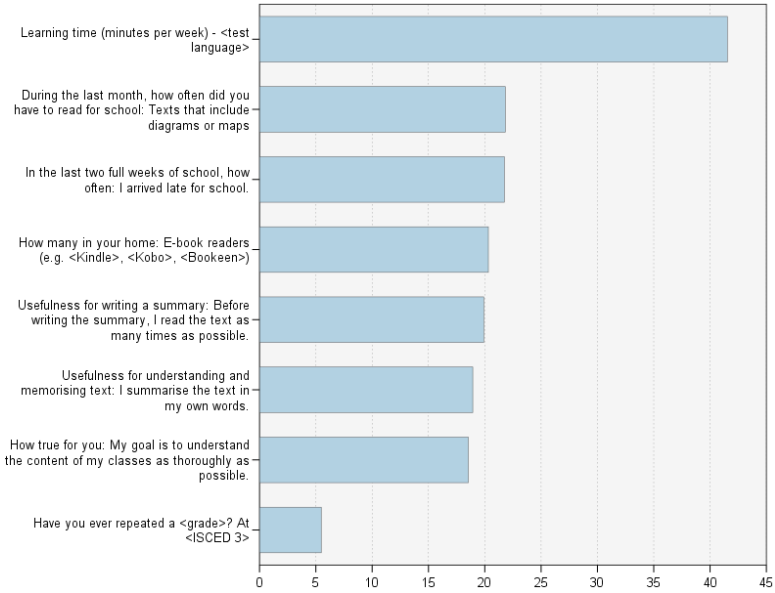
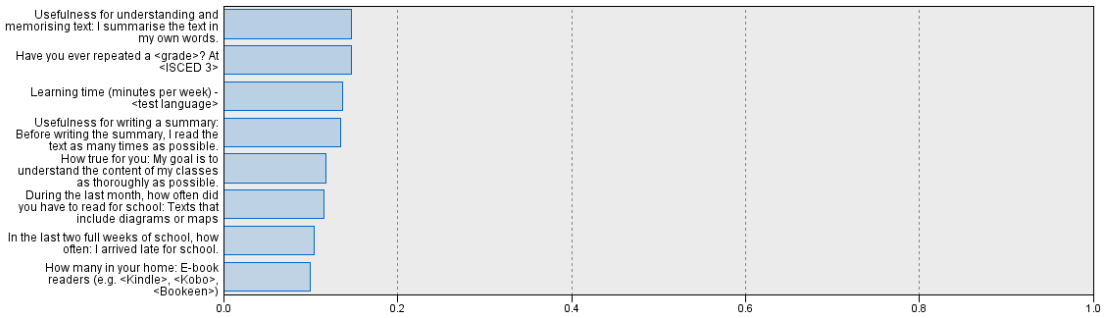


Figure4: Random Tree model results

ISSN: 2664-9640

The first two important classification variables of CHAID model output are Usefulness for understanding and memorizing text and Usefulness for writing a summary. The calculation results of this model highlight the impact of emotional cognition on students' reading literacy. The third grouping variable is "learning time", which reflects the important role of students' learning behavior and is the same as the results of CART algorithm.

The output results of the Random Tree model show that learning time is the most important classification variable, reflecting the importance of students' behavior. The second categorization variable is "how often did you have to read for school: Texts that include diagrams or maps"; The third important influence variable is "how often: I arrived late for school", which belongs to learning behavior representation.



*F*igure5: ANN model results

Different from the judgment results of other models, the model trained by ANN shows that the influence of multiple variables on students' reading literacy is not large, and the most important influencing factor is usefulness for understanding and... "Text" indicates the importance of understanding how to read. The second important factor is "have you ever repeated a grade", which is also very important for students to follow their history learning. The third essential classification factor is "learning time", which is also a method of evaluating students' learning.

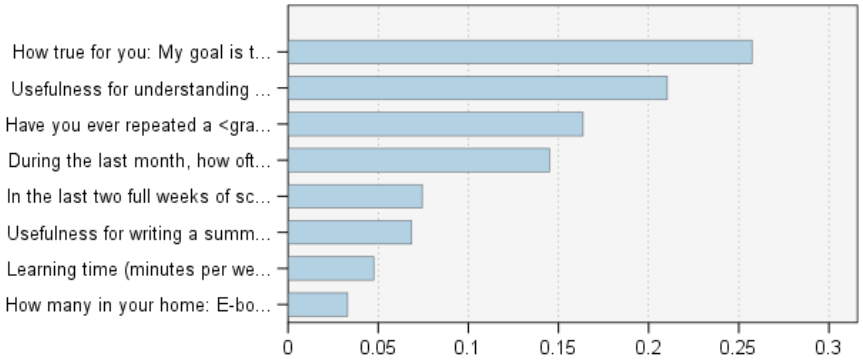


Figure6: LSVM model results

LSVM image reflects 8 effective variables, among which the most important classification variable is "How true for you: My goal is to understand the content of My classes as thoroughly as possible. The second grouping variable is "Usefulness for understanding and memorizing text: I summarize the text in my own words.", similar to the results of C5.0 and ANN algorithms; The third important grouping variable is " Have you ever repeated a <grade>?", reflecting the students' learning level of history.

In the result of DA, the first is "Usefulness for understanding and memorizing text", and the second important classification variable is "How true for you: My goal is to understand the content of My classes as thoroughly as possible." The third important categorical variable is "how often did you have to read for school".

ISSN: 2664-9640

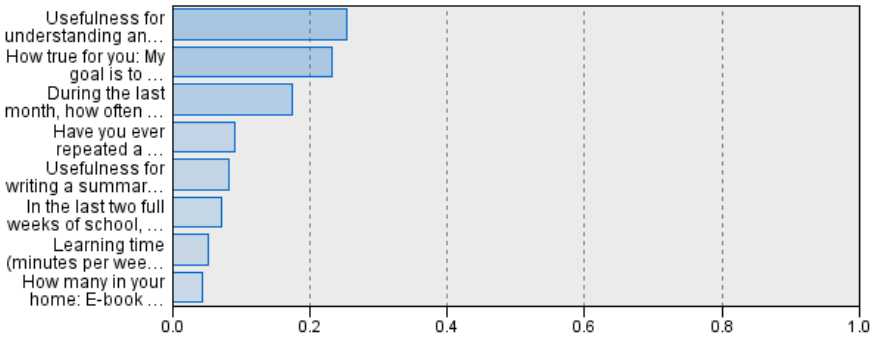


Figure7: DA model results

5. Conclusion

Data mining technology is more and more used in student performance prediction, which provides many constructive suggestions for teachers to improve their teaching level, and provides reasonable analysis methods for school management, thus improving the overall education level. However, most of the predictions of students' performance and ability focus on students' learning behavior, while this paper analyzes from the perspective of students' inner cognition and inner emotion. Based on students' inner cognition, this paper finds that the most important factor affecting reading literacy is “Usefulness for understanding and memorizing text: I summarize the text in my own words.” This cognitive variable, on the reading process pay more attention to summarize the students, his reading quality is higher. The least important variable is “How many in your home: e-book readers (e.g.)” about the number of e-books in the home, indicating that the existence of e-books greatly affects the students' reading level.

This paper also accurately classifies students with "reading literacy scores above the international average" and "reading literacy scores above the international average" through data mining algorithm. This paper uses four different classification algorithms, including three decision tree algorithms, support vector machine, random forest, neural network, discriminant analysis, all the classification algorithm test data set accuracy is more than 85%. This shows that data mining technology is very effective in predicting students' reading literacy by using PISA questionnaire data. It is hoped that this study will make effective contributions to education data mining and student reading literacy research.

In a word, the contribution of this study mainly includes two aspects. First, it makes use of students' emotional cognition and emotional state to predict reading literacy, which provides a new direction for future reading teaching. Secondly, the feasibility of applying data mining technology in PISA is verified by evaluating a variety of classification algorithms, and the decision tree classification model is worthy of model improvement, providing future research directions.

References

[1] Emmanuel N. Ogor: “Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques, Fourth Congress of Electronics”, Robotics and Automotive Mechanics Unrecognized Copyright Information, 78(2007), 354-359.

[2] Gökhan Aksu, Cem Oktay Güzeller: “Classification of PISA 2012 Mathematical Literacy Scores Using Decision-Tree Method: Turkey Sampling”, Education and Science. Vol. 41(2016), No. 185, p. 101-122.

[3] Farid Jauhari, Ahmad Afif Supianto. “Building student’s performance decision tree classifier using boosting algorithm”. Indonesian Journal of Electrical Engineering and Computer Science.Vol. 14(2019), No. 3, pp. 1298-1304.

ISSN: 2664-9640

[4] Xiang Hua, Yang Gongb, Chun Laib: The relationship between ICT and student literacy in mathematics, reading, and science across 44 countries: A multilevel analysis. Computers & Education, Vol.125 (2018),p. 1–13.
[5] Pisa 2018 Results Combined Executive Summaries Volume I, II & III, 2018.
[6] Student Questionnaire for PISA 2018 Main Survey Version, 2018.
[7] Pisa 2018 Assessment and Analytical Framework, OECD, 2019.
[8] Cao Hongjiang, Xie Jin: "Study on LSTM based prediction of academic performance and its influencing factors". Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition), Vol.22 (2020), p.90-100.
[9] Mustafa Agaoglu: "Predicting Instructor Performance Using Data Mining Techniques in Higher Education", Digital Object Identifier, Vol. 4(2016), p. 2169-3536.
[10] Christopher M. McLeod, N. David Pifer, Emily P. Plunkett: "Career expectations and optimistic updating biases in minor league baseball players". Journal of Vocational Behavior Vol.129 (2021), p.1-13.
[11] Suhas S Athani, Mayur N Banavasi: "Student Academic Performance and Social Behavior Predictor using Data Mining Techniques". International Conference on Computing, Communication and Automation (ICCCA2017). 2017.p.170-174.
[12] Xue wie: SPSS Modeler Data Mining Method and Application (Third Edition). Publishing House of Electronics Industry, 2020, p.32-167.
[13] Yannick Kiffen, Francesco Lelli, Omid Feyli. "A comparison between the Naïve Bayes and the C5.0 Decision Tree Algorithms for Predicting the Advice of the Student Enrollment Applications" [Online]. Information on: www.preprints.org.