Establishing an early warning system using machine learning algorithms to identify students at high risk of academic failure in online education

Xiangfeng Tan[1], Shumei Chen[1], Sumio Ohno[1], Hiroyuki Kameda[1] and Jinhua She[1]

[1] Tokyo University of Technology, 1404-1 Katakura, Hachioji, Tokyo 192-0982, JAPAN

Abstract. Monitoring student learning progress is important in online education. Identifying students at high risk of academic failure in the early stages of the course length and implementing timely teaching interventions can effectively improve learning outcomes. We analyzed the data log of the learning management system to extract students' learning activities and behaviors and converted them into variables. Using the data in the early stage of the course length, an early warning system based on a linear regression model was established to predict students' final exam scores and provide a reference for teachers to make decisions. The results show that the prediction accuracy of the EWS model is 81.1% when half of the course progress is completed. The results show the function of the early warning system in online education.

Keywords: Early warning system, learning management system, online education, machine Learning.

1 Introduction

With the rapid development of technology, school education has undergone radical changes in the past decades. Online education, as an innovative teaching mode, has gradually emerged and become an integral part of modern school education. School online education refers to a form of distance education realized through Internet technology and various digital means. Information technology has injected new vitality into school education, and the use and popularization of learning management systems (LMS) have greatly promoted the development of online education. With the development of online education, how to monitor student's learning progress and learning effects has become a major problem.

1.1 Online Education

The invention of the Internet has made online education more and more popular and started to create new teaching models. With the advancement of technology and the popularization of related equipment, the quality and acceptance of online education have been increasing [1]. The popularity of the Internet has greatly promoted the development of online education. The batch of online courses and virtual schools began to appear. Online education has provided a new experience for learners [2]. Online education plays an important role in today's colleges and universities [3].

For online education, there is a need for proper tools for online teaching and learning. The learning management system (LMS) is a software application used to manage educational programs [4]. Online education is usually operated by the use of LMS. LMS is widely used in various education fields [5]. LMS began to evolve with the popularity of the Internet. In 2000, the open-source LMS Moodle (Modular Object-Oriented Dynamic Learning Environment) was released and quickly became popular among educational institutions [6]. Since then, LMS has begun to evolve towards mobility and cloud computing, with platforms such as Canvas and Schoology supporting mobile apps that allow learners to access course content anytime, anywhere on their phones or tablets [7][8]. As computer technology advances and improves, the capabilities of LMS) are also increasing. Modern learning management systems are increasingly integrating artificial intelligence and data analysis to provide visual data summaries based on learner behavior data. [9]. During the COVID-19 epidemic, when schools in many countries of the world turned to online education due to the pandemic [10], the growing use of LMS is driving the popularity of online education and blended learning [11].

With the development of science and technology, online education in colleges and universities will continue to innovate the teaching experience of traditional face-to-face courses and create new features unique to online education. Introducing innovative methods such as playful teaching and autonomous learning to improve students' academic performance [12]. The future of online education will be a technological revolution [13].

1.2 Early Warning Systems

Compared to traditional school education face-to-face learning mode, Online education is conducted through an LMS, a large amount of student behavior data stored in the LMS system logs can be analyzed, making it possible to identify students who are at high risk of academic failure before the final exam. [14]. Early warning systems (EWS) in online learning are designed to identify students who are at risk of underachieving or dropping out of school, allowing teachers to intervene early and provide the necessary support [15]. LMS and other educational technologies are used to monitor student learning behavior and student academic progress [16]. Using EWS to identify students at high risk of academic failure and to monitor their behavior and academic progress is important for improving teaching outcomes [17]. Machine learning modeling and statistical analysis of online behavioral data and other forms of recorded data to identify students at high risk of failing is a common form of analysis used by EWS [18].

Using EWS can have a significant impact on improving academic performance in online education. Providing on-time support to students and answering their questions before they fall behind can improve their chances of success. It helps students have a smoother academic life by addressing issues that may lead to dropouts. Students can achieve better academic performance by identifying and addressing personal academic issues early [19].

The early warning system is very important in the field of online learning. The prediction of EWS has a very high degree of accuracy [20]. It predicts learning outcomes through data analysis and modeling and provides teachers with more convenient teaching reports for reference [21].

In this study, we analyzed log data collected from Moodle and constructed a prediction EWS model using linear regression to predict the final exam scores based on students' behavioral data at the halfway point in the course length. This prediction model can help teachers identify students at high risk of academic failure and intervene in advance.

2 Preliminaries

Establishing an early warning system requires collecting data, processing the data, and then selecting the appropriate model for modeling. Useful information is first extracted from the Moodle system data logs and converted into variables. After pre-processing, the data is analyzed and predictions are modeled using the right machine learning model. We introduce the methods used to establish an EWS and define the research objectives in this study.

2.1 Data pre-processing

Fig. 1 shows the detailed steps flow from extracting raw data from the Moodle system log to completing data pre-processing. The details of each step are as follows:

Determine Research Objectives. Determine specific goals for using Moodle log data based on student learning behaviors and course content design, and identify the general types of data required.

Data Extraction. Extract the log data from the Moodle system. Use data in Excel table format.

Data Classification. Clean and pre-process the extracted data. Populate the missing data. Remove outliers. Integrate and categorize the data. If data in one category is unusable, the implications of its use are reconsidered.

Data Variation. Summarize the data into various variables based on their characteristics.

Data Integration. Integrate all variables and remove variables with duplicate attributes.

Completion of data pre-processing. Data pre-processing is complete. Prepare for the next step of data analysis.

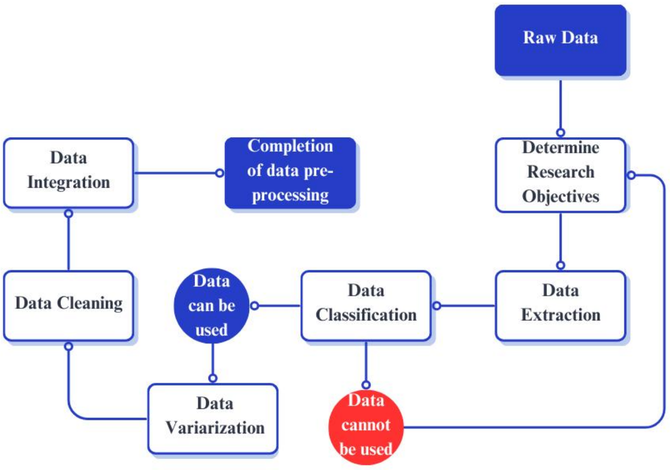


Fig. 1. Data pre-processing flow chart.

2.2 Data analysis

Analyze the pre-processed data set. Depending on the type of data, choose the appropriate analysis methods and understand the significance of the indicators of the results. In educational data mining, the following analysis methods are often used:

Chi-squared test. The Chi-squared test is used to analyze the relationship between two categorical variables. For example, the Chi-squared test is used for the researcher who wants to know the difference between two groups of students' preferences in online and offline teaching. If the analysis results show significance, then it means that there is a significant difference between the two groups.

Correlation analysis. Correlation analysis is used to find the relationship between quantitative data. Through analysis, it can be obtained whether there is a relationship and the correlation strength of the relationship. There are several correlation coefficients used to measure the degree of correlation. The most common one is the Pearson correlation coefficient. The strength of the relationship can be seen from the value of the correlation coefficient.

2.3 Machine learning

Machine learning (ML) is a technique for making predictions and decisions by analyzing and summarizing data and is often used in educational data analysis. Common machine learning methods can be divided into supervised learning and unsupervised learning.

Supervised learning. Supervised learning refers to training a model by training the input and output in the data set, and then using the model to predict new data. Linear regression is to establish a linear relationship between input variables and output variables. Continuous values can be predicted by a supervised learning model. Logistic regression uses a logistic function to map input variables to probability values between 0 and 1 and then classifies them based on the probability values. It can be used to predict discrete categories.

Unsupervised learning. Unsupervised learning refers to analyzing the characteristics of the data itself without labels. K-Means Clustering divides data points into K clusters through iterative optimization, making the data points in each cluster as similar as possible. It can be used for data clustering. Principal Component Analysis (PCA) projects high-dimensional data into low-dimensional space through a linear transformation while retaining the main information of the data. The characteristics of the data set can be extracted.

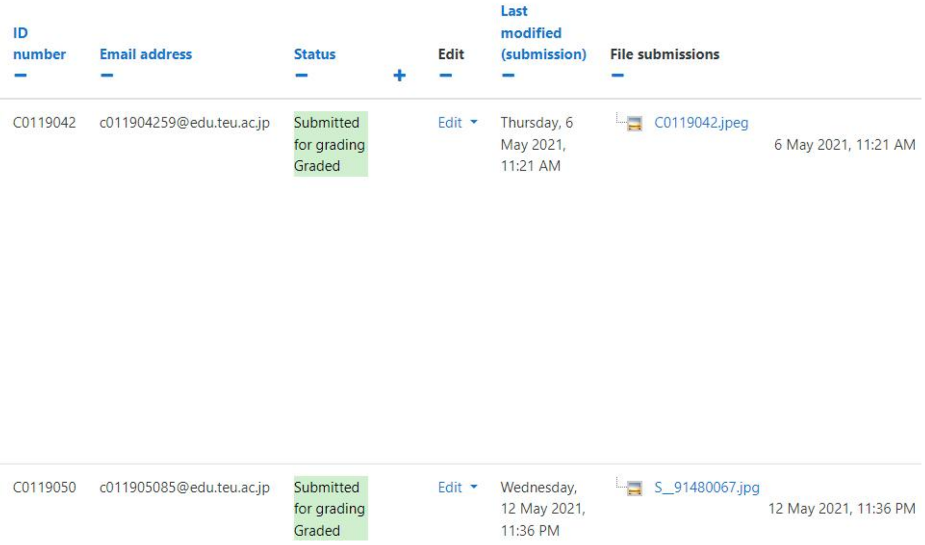
In this study, our study question is :

How to establish an effective EWS prediction model based on their early learning performance and online learning behaviors?

3 Data description

The sample of this study is from college students who are studying Chinese at a university in Tokyo, Japan. The course started in April 2021 and ended in July 2021. A total of 32 students completed the course. The course consists of 8 lessons, using an online flipped classroom teaching model, with listening, speaking, reading, and writing sessions. Students can freely choose their time to complete the course contents and conduct online teaching on ZOOM each week.

Fig. 2 shows the data that can be obtained from the Moodle background system. It includes student learning behavior data such as login times and homework completion times. We obtain data about students' online learning details from the system log.



ID number −	Email address −	Status −	+	Edit −	Last modified (submission) −	File submissions −
C0119042	c011904259@edu.teu.ac.jp	Submitted for grading Graded		Edit ▾	Thursday, 6 May 2021, 11:21 AM	C0119042.jpeg 6 May 2021, 11:21 AM
C0119050	c011905085@edu.teu.ac.jp	Submitted for grading Graded		Edit ▾	Wednesday, 12 May 2021, 11:36 PM	S__91480067.jpg 12 May 2021, 11:36 PM

Fig. 2. Moodle database guide.

After analyzing and extracting variables from the Moodle data log, we summarized five variables for subsequent analysis. These five variables represent the characteristics of learning behaviors. The following Table 1 gives a summary of all variables.

Table 1. Summary of all variables.

Variables	Description
$V_{\mathrm{i}\ (\mathrm{i}=1,\ \ldots,\ 4)}$	Video content views in Lesson 1-4
$O_{\mathrm{i}\ (\mathrm{i}=1,\ \ldots,\ 4)}$	Online days per week in Lesson 1-4
$R_{\mathrm{i}\ (\mathrm{i}=1,\ \ldots,\ 4)}$	Report completion times in Lesson 1-4
$T_{\mathrm{i}\ (\mathrm{i}=1,\ \ldots,\ 4)}$	Test completion times in Lesson 1-4
$H_{\mathrm{i}\ (\mathrm{i}=1,\ \ldots,\ 4)}$	Homework completion times in Lesson 1-4

Based on the extracted variables, we can explore whether there is a relationship between the students' final exam scores and these variables and whether the student's performance in the early period of the course can predict the students' final exam scores.

4 Methods

Fig. 3 shows the process of establishing an EWS model for this study. According to the students' online learning behavior, extract variables from the Moodle system log; perform data pre-processing on the variables; perform statistical analysis after pre-processing; select the appropriate machine learning method based on the conclusions drawn from the statistical analysis; and obtain the corresponding EWS model based on the selected machine learning method.

In this study, the total number of lessons in the sample is 8. We extracted the variables of students' learning behavior and performance from the 1st to the 4th lesson as independent variables. The final exam score at the end of the term was used as the dependent variable. An EWS using a linear regression model was constructed. The model can predict the final exam scores when the course is completed at half-length, providing a reference for teachers to intervene with students promptly.

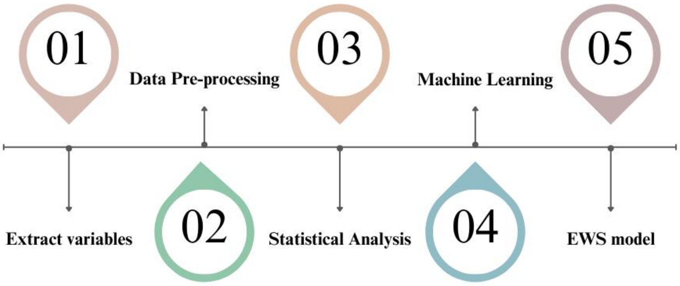


Fig. 3. Linear process of the study.

Data analysis and machine learning analysis were performed using R and SPSS. R and SPSS are used for statistical analysis, data visualization and data modeling based on machine learning algorithms. It can be used to perform statistical tests, regression analysis, time series analysis and machine learning algorithms.In evaluating the data, descriptive statistical methods were used to find differences between groups, and *t*-tests and one-way analysis of variance (ANOVA) were used for independent analysis. The Pearson correlation test was used to examine the level of correlation between two variables. The linear regression model was used to establish the EWS model. A multicollinearity test was performed to ensure that the basic assumptions were applied to the model.

5 Results

Results include descriptive statistics for the sample details and an EWS model constructed using a linear regression model.

5.1 Descriptive statistics results

Frequency refers to the number of times a numerical value representing a particular characteristic appears in the variable values. In this study, frequency analysis is used to determine the distribution of video views. We obtain the basic information of the research data at first; then, analyze each item one by one, and provide a detailed analysis of the options with larger selection proportions; finally, summarize the analysis results. Frequencies arranged by groups form a frequency distribution, which is used to explain the impact intensity of each group's score values on the overall score values.

The video teaching content is divided into three chapters. As shown in Table 2, from the distribution of V_1 data, most samples are "3.0," accounting for 96.88%. This means that most people have watched all the video content chapters of the first lesson. As the course progresses, the participation of students in video content study gradually decreases.

Table 2. Frequency analysis results**.**

Items	Categories	N	Percent (%)	Cumulative Percent (%)
V_1	1.0	1	3.13	3.13
	3.0	31	96.88	100.00
V_2	0.0	1	3.13	3.13
	2.0	4	12.50	15.63
	3.0	27	84.38	100.00
V_3	1.0	3	9.38	9.38
	2.0	7	21.88	31.25
	3.0	22	68.75	100.00
V_4	0.0	5	15.63	15.63
	1.0	2	6.25	21.88
	2.0	2	6.25	28.13
	3.0	23	71.88	100.00
Total		32	100.0	100.0

5.2 Pearson Correlation Analysis

Pearson correlation coefficient analysis is an analytical method to explore the degree and direction of linear correlation between two variables. Analysis can be performed when: both variables are continuous and come from the same sample group; there is a linear relationship between the two variables; and the two variables are bivariate normal distribution or approximate normal distribution.

We use correlation analysis to find the relationship between the number of online days per week and the completion of video teaching courses. As can be seen from Table 3, there is a positive correlation between O_1 and V_4; also, a positive correlation can be found between O_6 and V_4. There is no correlation between the other variables.

This shows that students who log in more days per week in the first and second classes of the course can maintain better learning concentration and persistence.

Table 3. Pearson Correlation analysis results.

	O_1	O_2	O_3	O_4
V_1	-0.077	0.173	0.127	-0.032
V_2	0.246	0.263	0.177	0.272
V_3	0.327	0.148	0.252	0.198
V_4	0.461**	0.416*	0.322	0.319

* $p < 0.05$ ** $p < 0.01$

5.3 The EWS regression model

Statistical modeling and regression analysis are the core implementation methods in machine learning, which are used to build prediction models to analyze and predict data. Linear regression is used to build a straight-line model. In this study, we used the modeling method of linear regression to build the EWS model using R with the students' final exam scores after completing the full course as the dependent variable and the students' learning behavior factors in the first half of the course as the independent variables. The model formula is: The model formula is:

$$Score = 61.743 - 11.392V_1 + 18.798V_2 + 2.309V_3 + 1.418V_4 - 1.727R_1 - 1.909R_2 + 5.178R_3 - 2.604R_4 + 14.976R_4 + 13.825T_2 + 6.315T_3 + 4.122T_4 - 4.359H_1 - 8.494H_2 + 2.013H_3 + 17.1154H_4 \quad (1)$$

As can be seen from Table 4, the model R-squared value is 0.811, which means that these variables explain 81.1% of the changes in Score. When the model is tested for *F*, it is found that the model passes the F test ($F = 4.022, p = 0.005 < 0.05$), Testing the multicollinearity of the model found that all *VIF* values in the model were less than 5, shows no collinearity problem; and the D-W value was under 2, which shows that there was no autocorrelation in the model and there was no correlation between the data.

By observing the model formula, we are able to determine the effect of each independent variable on the dependent variable by the positive and negative regression coefficients as well as the magnitude of the values. In this formula, V_2 has the largest positive regression coefficient, which shows the importance of this factor to the final scores. This model formula can be used to predict student performance.

Table 4. Summary table of linear regression model

	Unstandardized Coefficients		Standardized Coefficients	t	p
	B	Std. Error	*Beta*		
Constant	-61.743	58.620	-	-1.053	0.309
H_4	17.115	8.731	0.408	1.960	0.069
H_3	2.013	6.323	0.060	0.318	0.755
H_2	-8.494	6.625	-0.259	-1.282	0.219
H_1	-4.359	5.772	-0.104	-0.755	0.462
T_4	4.122	5.018	0.122	0.821	0.424
T_3	6.315	4.609	0.195	1.370	0.191
T_2	13.825	11.900	0.204	1.162	0.264
V_1	-11.392	7.860	-0.242	-1.449	0.168
V_2	18.798	4.151	0.687	4.529	0.000**
V_3	2.309	4.758	0.092	0.485	0.634
V_4	1.418	2.276	0.098	0.623	0.543
O_1	-1.727	1.285	-0.246	-1.344	0.199
O_2	-1.909	2.415	-0.136	-0.790	0.442
O_3	5.178	2.685	0.406	1.928	0.073
O_4	-2.604	2.512	-0.198	-1.036	0.316
R_4	14.976	8.629	0.302	1.735	0.103
R^2	0.811				
Adj R^2	0.609				
F	F (16,15) = 4.022, p = 0.005				
D-W	1.981				

Dependent Variable: *Score*
* $p < 0.05$ ** $p < 0.01$

6 Conclusion

This study proposes how to establish an EWS prediction model based on student log data. To address this problem, we used linear regression to build a prediction model and used variables summarized from log data in the first half of the course length to predict students' final exam scores, achieving a prediction accuracy of 81.1%. This study proves that early predictions can be made by extracting variables from students' learning behaviors.

Using the EWS model to predict and intervene in student behavior in the early stages of the course is of great benefit to improving students' academic performance.

It provides teachers with the opportunity to identify students at high risk of academic failure in online education classes and to promptly remind and intervene in these high-risk students, thereby improving the overall teaching effect. Having ample time to intervene with students halfway through the course can help teachers better arrange the course content progress and facilitate the decision-making process. In future studies, we plan to use innovative technologies such as VR software to create new teaching content, attract students to participate in the learning process, and improve learning participation and learning effects.

Acknowledgments

This work was supported in part by JSPS (Japan Society for the Promotion of Science) KAKENHI under Grant 22K00801.

References

1. Volery, T. and Lord, D.: Critical success factors in online education. International Journal of Educational Management **14**(5), 216-223 (2000)
2. Birchall, D.: Internet to facilitate global teaching and learning. Qualitative Market Research **3**(3), (2000)
3. Blaschke, L. M.: Heutagogy and lifelong learning: A review of pedagogical practice and self-determined learning. The International Review of Research in Open and Distributed Learning **13**(1), 56-71 (2012)
4. Harasim, L.: Shift happens: Online education as a new paradigm in learning. The Internet and higher education **3**(1-2), 41-61 (2000)
5. Cavus, N.: Selecting a learning management system (LMS) in developing countries: instructors' evaluation. Interactive Learning Environments **21**(5), 419-437 (2013)
6. Moodle Homepage, https://moodle.org, last accessed 2024/06/25
7. Canvas Homepage, https://canvas.ws/, last accessed 2024/06/25
8. Schoology Homepage, https://www.powerschool.com/, last accessed 2024/06/25
9. Ahmed, A. A. A., & Ganapathy, A.: Creation of automated content with embedded artificial intelligence: a study on learning management system for educational entrepreneurship. Academy of Entrepreneurship Journal **27**(3), 1-10 (2021)
10. Raza, S. A., Qazi, W., Khan, K. A., & Salam, J.: Social isolation and acceptance of the learning management system (LMS) in the time of COVID-19 pandemic: an expansion of the UTAUT model. Journal of Educational Computing Research **59**(2), 183-208 (2021)
11. Karma, I., Darma, I. K., & Santiana, I. M. M. A.: Blended learning is an educational innovation and solution during the COVID-19 pandemic. International research journal of engineering, IT & scientific research (2021)
12. Ebner, M., & Holzinger, A.: Successful implementation of user-centered game based learning in higher education: An example from civil engineering. Computers & education **49**(3), 873-890 (2007)
13. Elayyan, S.: The future of education according to the fourth industrial revolution. Journal of Educational Technology and Online Learning **4**(1), 23-30 (2021)
14. Macfadyen, L. P., & Dawson, S.: Mining LMS data to develop an “early warning system” for educators: A proof of concept. Computers & education **54**(2), 588-599 (2010)

15. Hu, Y. H., Lo, C. L., & Shih, S. P.: Developing early warning systems to predict students' online learning performance. Computers in Human Behavior **36**, 469-478 (2014)
16. Raffaghelli, J. E., Rodríguez, M. E., Guerrero-Roldán, A. E., & Bañeres, D.: Applying the UTAUT model to explain the students' acceptance of an early warning system in Higher Education. Computers & Education **182**, 104468 (2022)
17. McMahon, B. M., & Sembiante, S. F.: Re-envisioning the purpose of early warning systems: Shifting the mindset from student identification to meaningful prediction and intervention. Review of Education **8**(1), 266-301 (2020)
18. Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., ... & Khan, S. U.: Predicting at-risk students at different percentages of course length for early intervention using machine learning models. Ieee Access **9**, 7519-7539 (2021)
19. Osborne, J. B., & Lang, A. S.: Predictive Identification of At-Risk Students: Using Learning Management System Data. Journal of Postsecondary Student Success **2**(4), 108-126 (2023)
20. Mubarak, A. A., Cao, H., & Zhang, W.: Prediction of students' early dropout based on their interaction logs in online learning environment. Interactive Learning Environments **30**(8), 1414-1433 (2022)
21. Sletten, M. A., Tøge, A. G., & Malmberg-Heimonen, I.: Effects of an early warning system on student absence and completion in Norwegian upper secondary schools: a cluster-randomised study. Scandinavian Journal of Educational Research **67**(7), 1151-1165 (2023)