Analysis and Prediction of the Factors Influencing Students' Grades Based on Their Learning Behaviours in MOOCs

Ziyi Zhao, Fengxi Kang, Jing Wang, Binhui Chen,
Mingxuan Yang and Shaojie Qu

June 4, 2022

Analysis and Prediction of the Factors Influencing Students' Grades Based on Their Learning Behaviours in MOOCs

Ziyi Zhao
Information Center
Beijing Institute of Technology
Beijing, China
zhaoziyi@bit.edu.cn

Fengxi Kang
School of Mathematics and Statistics
Beijing Institute of Technology
Beijing, China
1120183117@bit.edu.cn

Jing Wang
School of Mathematics and Statistics
Beijing Institute of Technology
Beijing, China
1120192795@bit.edu.cn

Binhui Chen
School of Computer Science
Beijing Institute of Technology
Beijing, China
1120191040@bit.edu.cn

Mingxuan Yang
School of Xuteli
Beijing Institute of Technology
Beijing, China
1120181424@bit.edu.cn

Shaojie Qu*
Information Center
Beijing Institute of Technology
Beijing, China
qushaojie@bit.edu.cn

***Abstract*—The outbreak of COVID-19 brought new challenges to learning and teaching, and MOOCs (massive open online courses), as online distance learning platforms, provide new opportunities for teaching and learning activities. However, student learning efficiency is difficult to ensure in distance learning. Researchers have studied the relationship between students' grades and behaviours such as forum participation and video viewing; however, less research has been performed on students' submission behaviours. In this paper, we investigate the influence of learning attitudes reflected by students' submission behaviour and the trend in attitude change on grades. First, by studying students' submission behaviours, we identify new features that affect students' grades, such as students' resubmission behaviours. Second, we define positive attitudinal trends that students possess through student behaviour studies: more adequate code, more page viewing actions, and more aggressive submission details performance. Finally, we use the selected features to predict the students' performance. In the experiment, we predict student performance with an accuracy of 86.48%. This study will help teachers understand students' attitudes based on student behaviours and identify students who are struggling academically.**

Keywords*—*massive open online course, learning behaviour, learning initiative, performance prediction

I. Introduction

The development of MOOCs (massive open online courses) supports students learning through online distance learning methods; however, it is difficult to ensure students' learning efficiency in distance learning[1][2]. Researchers have studied students' behaviours and attitudes towards online learning and have tried to identify students who have academic difficulties[3][4]. Wu[5] and Liao[6] showed that learners who actively participate in online discussions and watch instructional videos in their entirety usually have satisfactory grades. In contrast, students who navigate the course with less action, barely interact with others, and plagiarize assignments usually receive lower grades[7][8]. Previous behavioural analyses have focused on course activities, after-school learning, and collaboration with others, with less research on students' behaviours in completing assignments. In this paper, we analyse students' learning attitudes and identify academically challenged students based on their behaviours on online programming assignments. The contributions of this paper are as follows: first, we find that certain features correspond to students' positive learning attitudes and significantly affect their performance. Second, we also introduce a period series to analyse the behavioural attitude trends of different students over different teaching weeks. Finally, we achieve an accuracy of 86.48% in predicting students' academic performance with the extracted features.

The article will be described in the following six sections. Section I introduces the research background of this study. Section II summarizes the findings of previous studies. Section III presents the research questions. Section IV discusses the data selection and the experimental method. Section V presents the analysis of the results of the research problem. Section VI is the conclusion.

II. Related Works

A. Students' behaviours in MOOCs

Researchers have conducted many studies on student behaviours in MOOCs. Liao[6] found that the main factors influencing the type of learners in a study of four courses on MOOCs were video viewing completion and complete sequences of different activities. Students with high grades typically had higher video completion and fewer completion sequences. Zhao[9] classified students into three categories of low, medium, and high well-being and clarified that students with high well-being prefer to explore courses by themselves rather than study for grades, thus achieving better grades in their studies. Researchers have also investigated the relationship between learning behaviours and learning attitudes. Chen[10] indicated that positive learning attitudes were effective in improving academic performance, which included actively engaging in course navigation as well as forum interaction behaviours. Onah[11] investigated the effect of learning attitudes on achievement in blended learning and showed that high-achieving groups of students are highly motivated, so they can effectively regulate their self-learning skills and improve their understanding through continuous postclass communication and group discussions.

B. Prediction

Regrading the selection of algorithms for achievement prediction models, researchers have studied various classification prediction algorithms, such as multiple regression models[12], neural networks[13], and decision tree[14], and conducted comparative analyses of model accuracy. Er[15] predicted the performance of 4358 learners by SVM and logistic regression, using the RMSE as a model performance assessment metric. The experimental results showed that SVM provided the optimal model performance, with an accuracy of 93%. Similarly, Huang[16] stated that if teachers need to predict students' academic performance by multiple variables, among multiple linear regression, MLP, and SVM, the SVM model should be chosen because it has the highest PAP among these four models. Injadat[17] and Migueis[18] compared SVM with ensemble learning and showed that ensemble learning was more accurate. Injadat[17] chose two-course stages of student performance, 20% and 50%, to predict the final student performance through SVM, linear regression, and bagging learning methods. The evaluation model metrics were the Gini index and p-value. The ensemble model was found to be more accurate in predicting both stages than any of the individual algorithms. Migueis[18] reported that by comparing algorithms such as SVM, naive bayes and random forest, it was found that random forest could achieve an accuracy of 96.1%. However, naive bayes had the worst prediction. However, Chen[19] proposed that naive bayes performed best in predicting the group of at-risk students in terms of achievement. The above related studies reveal that the results of the prediction models are influenced by the selection of features, assessment metrics, and study context. Therefore, the creation of prediction models requires the selection of suitable feature variables and assessment indicators. In this study, accuracy and recall will be used as the basis for model performance assessment; four prediction algorithms, namely, random forest, MLP, SVM, and naive bayes, will be compared; and the most appropriate algorithm will be selected as the final achievement prediction model for this study.

III. Research Questions

QUESTION A: What behaviours of students can have an impact on grades?

QUESTION B: How do student behaviours change over time?

QUESTION C: How can student grades be predicted through student behaviours?

IV. Data and Method

A. Data

The data for this study came from 1006 students in a *C Programming Language* course. These students came from more than 10 faculties, including the School of Computer Science, the School of Mathematics and Statistics, and the School of Management and Economics. The average age was 18 years old (the youngest age was 16 years old, and the oldest age was 21 years old), and students were concentrated in the freshman year, with a male to female ratio of 10:3. Additionally, the data of this study were anonymized so that students' personal privacy was ensured. A total of 13 submitted features about students were extracted as the focus of this study, as shown in Table I. The 13 features can be classified into three categories: Action, Detail, and Code.

TABLE I. FEATURES AND THEIR EXPLANATION

CATEGORY	FEATURE	EXPLANATION
ACTION	assign	Number of viewing programming assignment
	attempt	Number of attempts to answer questions before submission
	history	Number of viewing submission history
	reports_best	Number of viewing excellent program
	reports_detail	Number of viewing details of the submitted program, including completion time, codelines and codesize, and result of judgement
	result	Number of viewing programming results
	submit	Number of viewing code submission
	user report	Number of viewing submission reports, including the number of submitted programs, the number of passes, the number of successful compilation and so on
DETAIL	resubmitcountafterAC (resubmit)	Number of times the student resubmitted after "Accept"
	submit_avg_time (avg_time)	Average time for students to submit assignments
	submit_rank (rank)	Average submission ranking of students
CODE	sum_codesize (codesize)	Total code size submitted by students
	sum_codelines (codelines)	Total code lines submitted by students

Action refers to the eight submission — related behaviours in the submission log（assign, attempt, history, reports_best, reports_detail, result, submit, user_report）. The total number of students in each group is labelled as $LableUser[j_u]$, and the total number of actions is labelled as $LableAction[j_a]$. The average number of behaviours of students in each group is calculated by equation (1). It can provide data support for the subsequent establishment of an analysis model.

$$M_{action} = LableUser[j_u]/LableAction[j_a] \quad (1)$$

Detail refers to the behaviours of students at the time of submission, which is more like the unconscious behaviours of students. *Submit_avg_time* indicates the average completion time of students in a certain programming topic; this feature can reflect the seriousness of students regarding the completion of the assignment. *submit_rank* represents the submission rank of students. By extracting the earliest submission record of students for each programming topic, we can determine the submission rank of students in a certain programming topic, which can reflect their active submission behaviour. *resubmitcountafterAC* represents students' resubmission behaviour after "Accept". There are 8 types of submission results for programming questions, including AC(Accept), WA (Wrong Answer), and CE (Compile Error). However, only when the submission judge result is AC is the programming topic passed. We look up the submission sequence corresponding to the submission time when the student passed the program and compare it with the total number of submissions of the student to determine whether the student has the behaviour of resubmitting after AC.

Code refers to the total amount of code submitted by the student when the first judge result of the programming assignment is AC. C_{size} is defined as the total code size submitted by students, C_{line} is defined as the total number of lines of code submitted by students, $U_x(1 \leq x \leq 1006)$ is defined as the number of students, and $P_y (1 \le y \le 72)$ is defined as the number of programs. An analytical model is constructed by calculating the average code size and the average number of code lines for students in each grade band to explore the correlation between the amount of code and grades.

$$\bar{X}_{size} = \frac{1}{U_X} \sum_{x=1,\ y=1}^{n} (C_{size_{U_x P_y}}) \quad (2)$$

$$\bar{X}_{line} = \frac{1}{U_X} \sum_{x=1,\ y=1}^{n} (C_{line_{U_x P_y}}) \quad (3)$$

B. Method

1) METHOD OF QUESTION A

Students' different behaviours will affect their performance. In this part, features extracted from the data processing part are used to establish the RFECV feature selection model based on random forest. After the fitting of random forest feature attributes, feature attributes are divided according to the importance degree, and the weight value will be given. The weight value represents the influence of each attribute on the label attribute. The larger the weight value is, the greater the influence is, and the smaller the weight value is, the smaller the influence is. Therefore, the weight value also represents the importance of each attribute to the accuracy of prediction. As shown in formula (4), $Weight_i$ is the weight of the i^{th}features and sum of all features' weight adds up to 1[23]. In addition, this study establishes the correlation between 13 features and students' grades and explore the behavioural differences of students with different grades.

$$\sum_{i=1}^{n} Weight_i = 1 \quad (4)$$

According to the influence of students' behaviour on performance, we define the behaviour that has a positive impact on performance as a positive learning attitude, and the behaviour that has a negative impact on performance as a negative learning attitude. Positive learning attitudes include: more adequate code and more page viewing actions, such as reviewing past assignments, and reviewing excellent programming programs; more aggressive submission behaviour, such as earlier assignment submission, repeated submission after "Accept".

2) METHOD OF QUESTION B

The change trend of different students' behaviour is different. We can distinguish students with different grades through the change trend of behaviour. The opening time of programming questions is mainly from week 4 to week 19 of the teaching week. The average number of repeat submitters in the 16 teaching weeks for groups of students with grades below 50 and above 80 is calculated. The learning attitudes of students in different grade groups throughout the semester are judged. Through the comparative analysis of the average number of students resubmitting in 16 teaching weeks, the influence of students' attitude change trend on their grades is explored.

3) METHOD OF QUESTION C

In this part, four models, naive bayes[20], MLP[21], SVM[22], and random forest[23], are developed to predict students' performance by extracting 13 behavioural features of students. The final results are compared and studied to find the most appropriate model for predicting performance. Since the median value of the student score band is 82, a score of 80 can minimize the difference in data volume between the left and right sides in data splitting. Moreover, this study also trichotomizes the data by 75 and 90 grades to further explore the impact of student behaviour on student performance. The prediction performance is judged by the following metrics: accuracy and recall. Accuracy represents the accuracy of the model, and recall represents the percentage of records with positive predictions that are correct.

V. RESULTS

A. Results of question A:

As shown in Fig. 1, we find that each feature has different feature importance in ranking on grades using RFECV feature selection on 13 features. Of the 13 features that had the most impact on performance, submit_rank (rank) scores 12%. The effect of user_report on performance is only 0.3%.

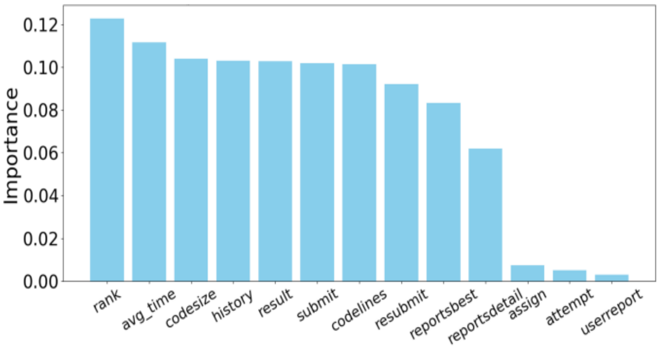

Fig. 1 Comparison of the effects of each feature

Then, we analyse the features by categories. Fig. 2 shows that students' performance is positively correlated with the average number of actions among the 8 behaviours involved. Especially in the behaviour of *reports_best* (Fig. 3), the frequency of this behaviour tends to be 0 for students with grades below 20. Therefore, it is clear that students who have higher grades pay more attention to their past assignment performance and have more interest in viewing the excellent program examples, which indicates a positive attitude towards the course.

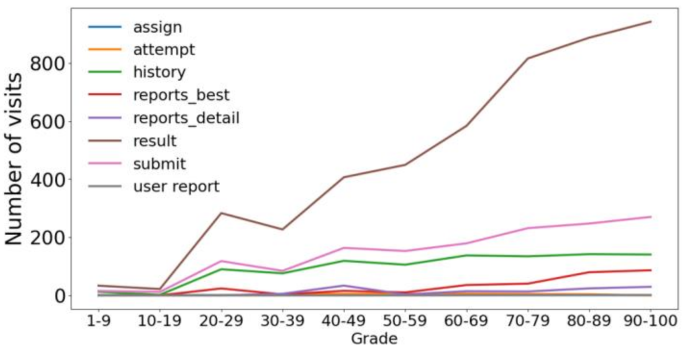

Fig. 2 Relationship between action and grade (8 actions)

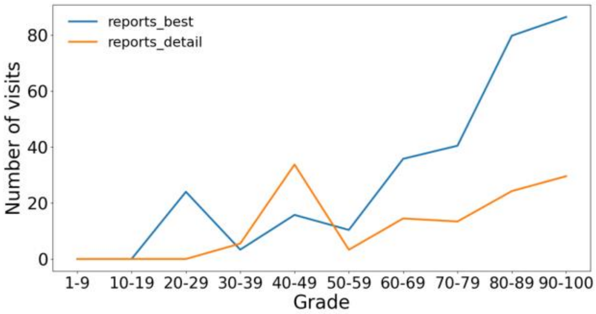

Fig. 3 Relationship between report_best times and report_detail times and grade

As shown in Figs. 4, 5, and 6, the number of resubmissions and the average submission time are positively correlated with the student's grades; that is, students with higher grades are more willing to submit repeatedly and spend more time studying the problems, thus demonstrating their desire to explore. From the submission ranking, grades and the submission ranking show an inverse trend, indicating that students with higher grades prefer to submit programming assignments earlier rather than submit the work in a hurry before the deadline. The earlier submission means that students have more time to prepare, and the rest of the time could be used to find new solutions or preview the new curriculum, also illustrating how students attach importance to assignments.

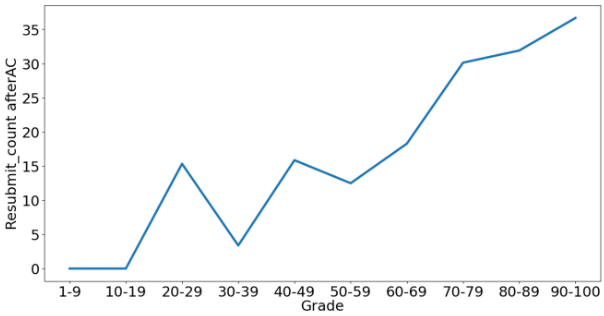

Fig. 4 Relationship between resubmit and grade

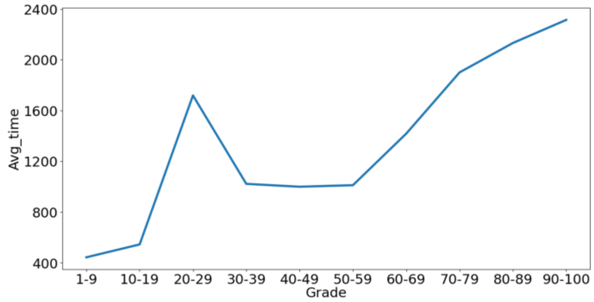

Fig. 5 Relationship between average time and grade

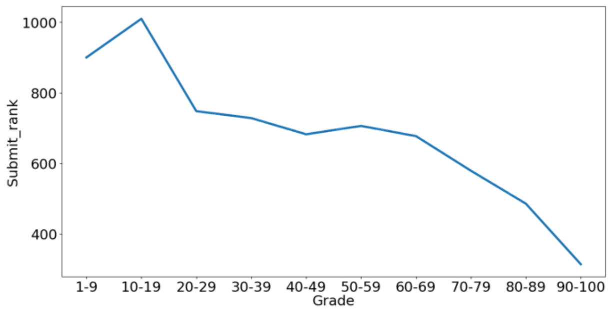

Fig. 6 Relationship between submission rank and grade

In the analysis of the Code class, we select 66,000 code records of students passing the first time from 191,180 student submission records to analyse the difference in the amount of code in different grades. As shown in Fig. 7, codelines and codesize are positively correlated with grades as grades increase gradually. This means that students with higher grades submit more code. To some extent, the amount of code reflects students' understanding of the programming courses. Although brief codes are more efficient and take less time, they are more demanding for students' logical capabilities. However, most of the students choosing this course are freshmen and do not know much about programming before having this course. Therefore, a larger amount of code could enhance the readability of code and thus help them establish better logic of programming.

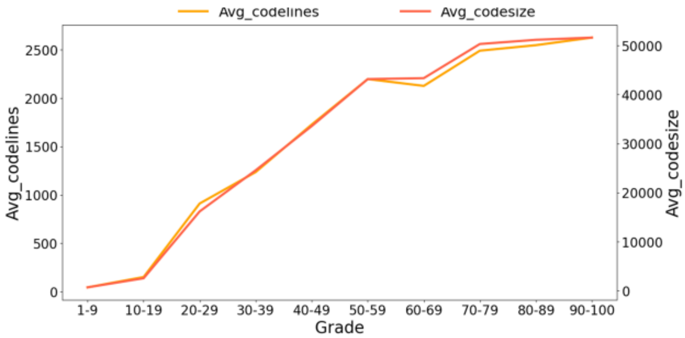

Fig. 7 Relationship between code (size & lines) and grade

Through the above analysis, students' learning attitude can be judged according to the three types of behaviours: Action, Detail and Code. Positive learning attitude includes: 1) More viewing actions (assign, attempt, history, reports_best, reports_detail, result, submit, user_report); 2) More aggressive submission details, including earlier submission of assignment and more repeated submission behaviour after "Accept"; 3) More accurate code, including more codelines and codesize.

B. *Results of question B:*

Through the analysis of the two trend axes (Fig. 8), it can be seen that the number of repeat submissions of students with more than 80 points fluctuates with the change in teaching week; however, it basically tends to be stable. For the group of students with grades below 50, the number of students who choose to submit repeated that decreases with time, and the whole curve shows a downwards trend. Therefore, for the group of students with high grades, their learning attitudes will not change greatly with time. However, for the group of students with low grades, their learning attitudes will gradually become less positive over time.

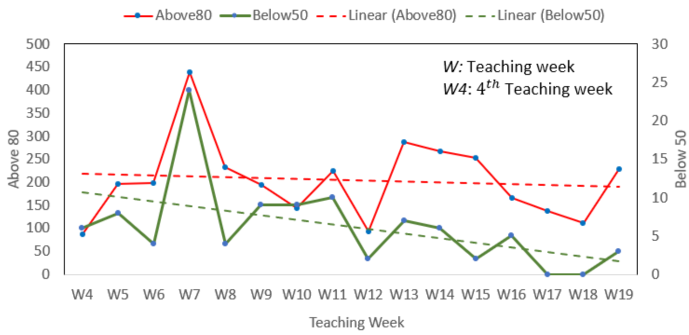

Fig. 8 Relationship between number of re-submitters and teaching week

C. Results of question C:

We use relevant features to predict students' performance through relevant methods of machine learning. Through the analysis and comparison of the four algorithms in Table II, it is found that the accuracy and recall of random forest are better than those of the other algorithms. Among the binary classification method with 80 as the classification standard, the accuracy of random forest is 86.48%, and the recall of prediction for students with fewer than 80 points is 81.28%. For students above 80 points, the prediction recall is 90.79%. For the three-way classification, although random forest has a higher accuracy of 74.06%, its performance is unsatisfactory compared with that of the binary classification method. In the three-way classification, the number of students is divided into three equal quantity according to the grades of 0-75,75-90 and 90-100, which can ensure the balance of data quantity, nevertheless, because the average behaviour times of students with scores of 75-90 and 90-100 are not much different, the learning effect of the model is not ideal and the accuracy is not satisfied. Moreover, this part also introduced the five-fold cross validation model to verify and analyse the accuracy of the model, and the result indicated that the accuracy of random forest was the highest, reaching 80.00% +/- 8.66%.

TABLE II. Binary Classification

Binary Classification				
Classification		Accuracy	Recall	Cross Validation
SVM	below 80	78.23%	75.77%	77.13% +/- 3.13%
	above 80		80.25%	
Naive Bayes	below 80	75.55%	77.97%	74.84% +/- 5.11%
	above 80		73.55%	
Random forest	below 80	**86.48%**	81.28%	80.00% +/- 8.66%
	above 80		90.79%	
MLP	below 80	76.64%	83.26%	77.43% +/- 4.32%
	above 80		71.20%	

TABLE III. Three-way Classification

Three-way Classification				
Classification		Accuracy	Recall	Cross Validation
SVM	below 75	60.23%	63.11%	60.03% +/- 1.48%
	between 75 and 90		56.81%	
	above 90		61.59%	
Naive Bayes	below 75	60.14%	82.01%	59.34% +/- 5.04%
	between 75 and 90		35.48%	
	above 90		68.51%	
Random forest	below 75	**74.06%**	68.90%	68.29% +/- 7.05%
	between 75 and 90		82.78%	
	above 90		68.17%	
MLP	below 75	62.62%	73.48%	59.44% +/- 4.52%
	between 75 and 90		50.13%	
	above 90		67.13%	

VI. CONCLUSION

This paper studies the relationship between student behaviour and grade on MOOC platform. By analysing the influence of 1006 students' behaviour in *C language programming* on their grades, this paper identified three behavioural features of students with positive attitude: high behaviour frequency, excellent performance in assignment submission behaviour details, and a large number of code specifications submitted. In addition, the resubmission behaviour of students after "Accept" is analysed in the period dimension, and the analysis finds that for students with high grades, their repeated submission times did not change significantly over the whole semester, and their learning attitudes remained stable. However, for students with low grades, over time, their repeated submission times decreased, and their learning enthusiasm gradually decreased. Based on the above 13 students' behavioural features, the students' academic performance is predicted and analysed. The experimental results show that the accuracy of random forest is the highest, reaching 86.48%.

Although this paper studies the factors of programming assignment submission that affect students' performance, the accuracy and recall of the prediction model can be further improved. We will study more students' submission behaviours in the future, including the detailed analysis of the compiled codes submitted, so as to improve the prediction effect of the overall model.

REFERENCE

[1] C. Coman, LG, Țîru, L .Meseșan-Schmitz, C. Stanci, MC . Bularca, "Online Teaching and Learning in Higher Education during the Coronavirus Pandemic: Students' Perspective," Sustainability. 2020, 12(24):10367.

[2] OF. Ayanbode, A. Fagbe, R. Owolabi, S. Oladipo, OR. Ewulo, AYMA. Islam, "Students' interactions, satisfaction and perceived progress in an online class: Empirical evidence from Babcock university Nigeria," Cogent Education 9:1,2022.

[3] Y. Meier, J. Xu, O. Atan and M. van der Schaar, "Predicting Grades," in IEEE Transactions on Signal Processing, vol. 64, no. 4, pp. 959-972, Feb.15, 2016.

[4] D. Hooshyar and Y. Yang, "Predicting Course Grade through Comprehensive Modelling of Students' Learning Behavioral Pattern," Complexity, vol. 2021, Article ID 7463631, 12 pages, 2021.

[5] M. Wu, H. Zhao, X. Yan, Y. Guo and K. Wang, "Student Achievement Analysis and Prediction Based on the Whole Learning Process," 2020 15th International Conference on Computer Science & Education (ICCSE), 2020, pp.123-128.

[6] P. Liao, J. Xu, S. Gong, W. Liu and Y. Yi, "Clustering Analysis of Learners' Watching Sequences on MOOC Videos," 2021 16th International Conference on Computer Science & Education (ICCSE), 2021, pp. 111-116.

[7] H. Wang, X. Hao, W. Jiao and X. Jia, "Causal Association Analysis Algorithm for MOOC Learning Behavior and Learning Effect," 2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science

and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), 2016, pp. 202-206.

[8] S, EI. and R. Ghada, “Understanding How Learners Use Massive Open Online Courses and Why They Drop Out: Thematic Analysis of an Interview Study in a Developing Country,” Journal of Educational Computing Research, 2016:0735633116681302.

[9] C. Zhao and C. Mi, “A Study on the Differences of Attitude Learning and Learning Behavior Sequences for University Computer MOOC+ SPOC Course,” 2020 15th International Conference on Computer Science & Education (ICCSE), 2020, pp. 365-369.

[10] P. Chen and Y. Chen, “Massive Open Online Course Study Group: Interaction Patterns in Face-to-Face and Online (Facebook) Discussions,” Frontiers in Psychology, vol. 12, 2022, pp.1664-1078.

[11] D.F.O. Onah, E.L.L. Pang, and J.E. Sinclair, “Cognitive optimism of distinctive initiatives to foster self-directed and self-regulated learning skills: A comparative analysis of conventional and blended-learning in undergraduate studies,” *Educ Inf Technol* 25, pp.4365–4380, 2020.

[12] B. Rienties and L. Toetenel, “The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules,” Computers in Human Behavior, vol. 60, 2016, pp. 333-341, 0747-5632.

[13] I. Lykourentzou, I. Giannoukos, G. Mpardis, V. Nikolopoulos and V. Loumos, “Early and dynamic student achievement prediction in e-learning courses using neural networks,” Journal of the American Society for Information Science & Technology, 2014, 60(2):372-380.

[14] N. Fang and JG. Lu, “A Decision Tree Approach to Predictive Modeling of Student Performance in Engineering Dynamics,” INT J ENG EDUC, vol. 26, 2010, pp. 87-95.

[15] E. Er, E. Gómez-Sánchez, M L. Bote-Lorenzo, Y. Dimitriadis and JI. Asensio-Pérez, “Generating actionable predictions regarding MOOC learners’ engagement in peer reviews,” *Behaviour & Information Technology* 39:12, 2020, pp. 1356-1373.

[16] S. Huang and N. Fang, “Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models,” Computers & Education, vol. 61, 2013, pp. 133-145, 0360-1315.

[17] M. Injadat, A. Moubayed, AB. Nassif and A. Shami, “Multi-split optimized bagging ensemble model selection for multi-class educational data mining,” *Appl Intell* 50, pp.4506–4528, 2020.

[18] VL. Migueis, A. Freitas, PJV. Garcia, and A.Silva, “Early segmentation of students according to their academic performance: A predictive modelling approach,” Decision Support Systems 115, pp. 36-51, November 2018.

[19] H. M. Chen, B. A. Nguyen, Y. X. Yan and C. R. Dow, “Analysis of Learning Behavior in an Automated Programming Assessment Environment: A Code Quality Perspective,” in IEEE Access, vol. 8, 2020, pp. 167341-167354, 2169-3536.

[20] K.Nigam, AK. Mccallum, Sebastian, T and T. Mitchell, “Text Classification from Labeled and Unlabeled Documents using EM,” Machine Learning, vol. 39, pp. 103–134, 2000.

[21] MW. Gardner and SR. Dorling, “Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences,” Atmospheric Environment, vol. 32, pp. 2627-2636, Aug.1, 1998.

[22] C. Cortes and V.Vapnik, “Support Vector Networks,” Machine Learning, vol. 20, pp. 273-297,1995.

[23] L. Breiman, “Random Forests,” Machine Learning, vol. 45, pp. 5–32, 2001.