Using Cluster Analysis to Explore Students' Learning Time Preference in Online Education

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

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

Abstract. The rise of online education provides convenience for students to arrange their learning time flexibly. Due to the large number of students who study in the evening, it is meaningful to explore the impact of late-night studying on academic performance. In this study, we used statistical analysis and machine learning to study the distribution of learning time periods for students with different study habits, and the impact of learning time periods on learning attitudes. The results of the study show that there are differences in the learning time period preferences of students. Students who prefer to study during the day perform better in academic performance.

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

1 Introduction

In recent decades, with the wide application and continuous innovation of computer information technology, the Internet has changed the way of life of human beings. With the development of science and technology, university education has also undergone significant changes due to the advancement of the teaching environment and the innovation of teaching methods. Online education has had a profound impact on teaching methods [1]. With the rapid advancement of information technology, educational content and methods have continued to develop, continuously reducing the difficulty of learning [2]. With the widespread use of home internet-connected electronic devices, online education provides many new learning options compared to traditional face-to-face education [3]. Current online education allows students to access learning materials and complete their studies anytime and anywhere [4]. For students, online education permits them to maximize their learning efficiency by allowing them the flexibility to arrange the time and content of their studies without the constraints of a fixed schedule [5].

1.1 Sleep quality and study performance

Studying late at night has always been an inevitable part of students' lives, especially during the revision days before final examinations [6]. The state of human mental activity is significantly different during the day and night [7]. Several studies have investigated the impact of circadian rhythms on human mental states. In a survey of college students, a negative correlation was found between sleep time length and perceived stress for most participants, and increased stress during the day was associated with shorter sleep time length at night [8]. In a study on the relationship between the quality of sleep at night and next-day mood among adolescents in the United States, researchers found that improving the quality of sleep at night can help improve adolescents' sense of well-being and reduce negative emotions the next day [9]. Good sleep quality at night helps improve personal cognitive performance [10]. According to the results of the literature survey, there is a correlation between insufficient sleep and students' academic performance [11]. The effects of insufficient sleep are mainly reflected in higher-level cognitive functions such as attention, memory, and problem-solving; which harms learning ability and academic performance [12]. Using personal electronic devices at night is associated with reduced total sleep time length [13].

In online learning environments where students can study flexibly at any time, many students are choosing to study at night. The pros and cons of late-night studying and its potential impact on student's physical and mental health are common research topics in studies related to the effects of late-night studying on students' performance. Current research mostly uses questionnaires to investigate students' study time preferences. Our research provides a fundamentally fresh perspective on understanding the issue by using data analysis for LMS logs and by using machine learning algorithms to tap into students' study time preferences. Our research explores students' preferred study time of the day in online education and the impact of study time on students' completion of learning content.Our study explored students' preferred time period of the day to study in online education and the impact of the time period of study on students' completion of the learning content.

1.2 Educational Data Mining

With the rapid development of online education technology and the widely used of learning management systems (LMSs), online education platforms provide more and more detailed raw data for data analysis. Educational data mining (EDM) has become an important tool for teachers to improve course content design. [14]. EDM is widely used in online education, including analyzing and visualizing test scores, evaluating students' learning behaviors, predicting students' academic performance, providing personalized learning guidance to students, and providing students' summary evaluation reports to teachers [15]. After the Covid-19 pandemic, university education around the world has changed radically [16]. The rapid popularization of online education in recent years has brought about an urgent need to improve the quality of online education. EDM has been used more and more frequently in the online education field. Machine learning has been widely used in educational data analysis. EDM uses a variety of machine learning algorithms for analysis, such as neural networks, decision trees, and cluster analysis [17]. By analyzing students' learning behavior data, cluster analysis is used to categorize students and reveal differences in learning patterns among groups of students [18]. Based on the results of the analysis, targeted teaching and management can be carried out for groups of students with different characteristics. Improvement of current weaknesses in learning and enhancement of students' learning efficiency.

This study explores the practical usefulness of students' characteristic behavioral data using EDM. The study questions are as follows:

1. Is there a correlation between the study habits of students who study during the time period of midnight (i.e. 24:00 to 4:00) and learning motivation?
2. What kind of information can we get by analyzing the classification results of cluster analysis?

To investigate these study questions, this study collected data on students' online behaviors provided by the Moodle system log data under the online flipped classroom learning model. Our study aims to explore and analyze the data to investigate the differences in individual student preferences for learning time periods, and the combined effect of students' flexibly selected learning time periods on their academic performance in online learning to provide more scientific guidance for course improvement.

2 Online teaching system and analysis methods

Most students study during the day and night, but a few students study during late nights and early mornings. Online video learning contents in our study is the central part of the course considered in this study. The overall learning motivation of students can be assessed by whether they can complete and watch all video chapters. We can use Pearson correlation analysis to analyze the correlation. The results can be used to assess whether studying late at night affects their learning attitudes.

In this study, we obtained the raw data for analysis and pre-processed the data through a well-designed course session. The data was analyzed by selecting appropriate analytical methods.

2.1 Structure of teaching system

As can be seen in Fig. 1, the sequence of completion of the learning course sequence in this study. The course was to be taught through the Moodle learning platform online. The course structure is based on the online flipped classroom model. In the online flipped classroom model, students are required to learn the course contents through self-learning, and then communicate with the teacher and classmates through online real-time communication software.

The course sessions are organized as follows:

Video learning materials. At the beginning of the lesson, students need to watch the video learning materials. The video learning materials consist of three chapters, and the length of each chapter is 5 minutes.

Report. After they watched the video learning materials, students were required to write their own learning experience through a short report, including what they had already known and what they didn't understand from the video learning materials.

Test. Based on the knowledge points learned in the video learning materials before, students need to complete an online time-limited test.

Extra learning content. The extra learning content is optional and it provides more relevant learning materials for students who are well-motivated in learning.

Homework. Students need to complete their homework by handwriting and upload homework to the Moodle platform within a certain deadline.

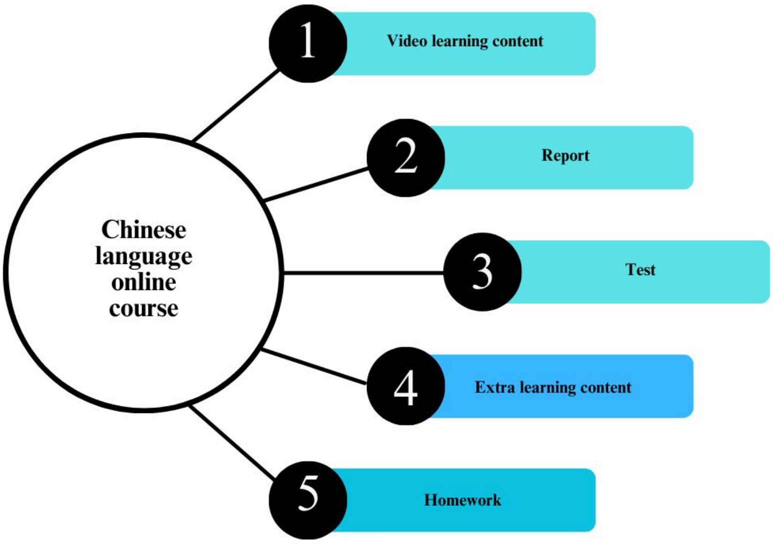


Fig. 1. Course learning sequence.

2.2 Data processing

Systematic data logs from Moodle recount students' basic activity status and online learning behaviors each time they are logged in. The data are categorized, and extracted into feature variables according to the demands of the study.

2.3 Statistical analysis

Based on the understanding of the meaning of the variables extracted from the data log, we visualize the data to better understand the changing trend of the data. Data visualization and descriptive analysis of data can reveal correlations between data and overall distribution. According to the study questions, select the corresponding variables for analysis, and analyze with the right analysis methods.

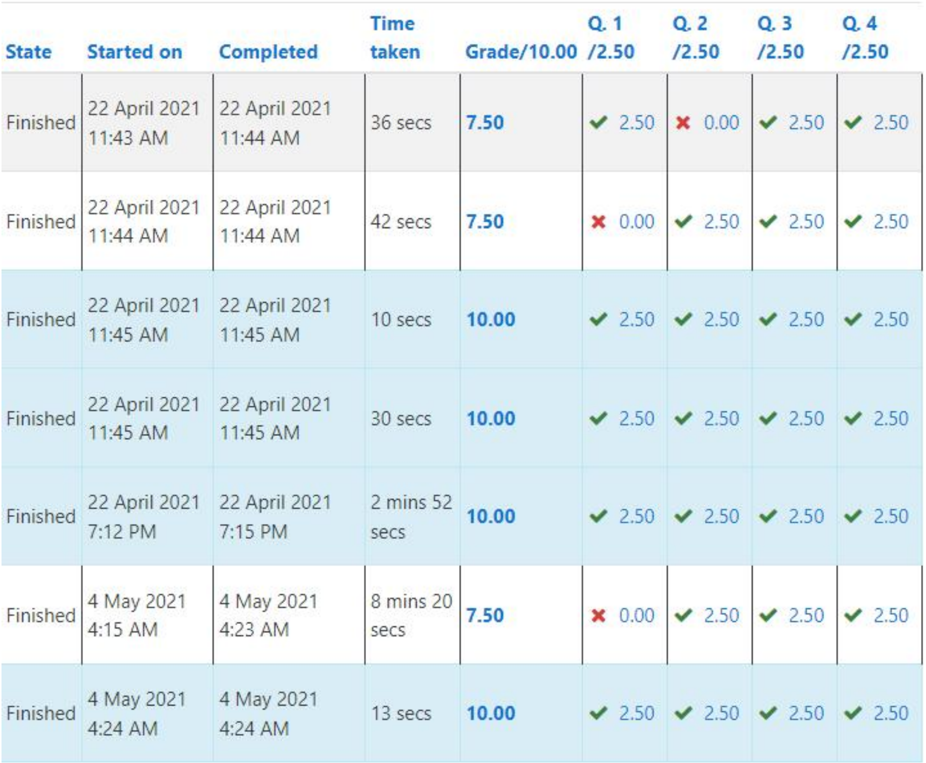
2.4 Machine learning

Machine learning (ML) has important usability in EDM. With its ability to help teachers better understand students' learning behavior and learning effectiveness, it improves the quality of course design. Machine learning is divided into two categories: supervised learning and unsupervised learning. Unsupervised learning is the analysis and modeling of data without labels. The researchers do not need to annotate the data with labels beforehand. Unsupervised learning discovers hidden relationships and structures in data. Using unsupervised learning in the education field helps teachers to better identify hidden issues and provide personalized teaching methods and special support appropriate for different groups of students. Informing course curriculum improvement. Before performing cluster analysis, we require choosing the appropriate cluster algorithm. In this study, K-prototype cluster algorithm is used for analysis. K-prototype cluster algorithm combines K-means and K-modes cluster algorithms. It is used for raw data sets that contain numerical and categorical variables. The K-prototype cluster algorithm is initialize the cluster centroids, containing the values of numerical and categorical variables. Calculate the distance of each data point from the cluster centroids. Assign each data point to the nearest cluster based on the calculated distances.

As students may have other things to prioritize during the day and night, such as assignments in other subjects, part-time jobs to earn a living, etc., online learning time is highly influenced by other real-life activities. The learning time factor is highly random. It is very difficult to determine whether students prefer to study during the day or at night. Unsupervised learning using machine learning is effective in solving this problem without the marked labels. In terms of analytical tool selection, R and SPSS were used for data analysis and machine learning classification. Descriptive data and data visualization were used to show data differences. Pearson correlation analysis was used to find the correlation between two variables. Cluster analysis has been used to classify patterns of learning behavior in students. Since the results of the cluster analysis required us to find the differences between groups ourselves, Analysis of variance (ANOVA) was used to explore the significance of the classification results. The chi-square test was also used for analysis.

3 Data collection

The sample for this study was 99 students enrolled in a Chinese learning course at a university in Tokyo, Japan. Data were collected from online learning tracks students left after completing various learning contents on Moodle. The course is taught by an online flipped classroom model. The course was completed in the first semester of 2021. As can be seen in Fig. 2, the interface of the Moodle teaching backend management system. By summarizing and generalizing the characteristics of the Moodle system log data, we extracted variables from it for analysis. Combined with variables extracted from data obtained from other data sources, this constitutes the full set of variables used in this study.



State	Started on	Completed	Time taken	Grade/10.00	Q. 1 /2.50	Q. 2 /2.50	Q. 3 /2.50	Q. 4 /2.50
Finished	22 April 2021 11:43 AM	22 April 2021 11:44 AM	36 secs	7.50	✓ 2.50	✗ 0.00	✓ 2.50	✓ 2.50
Finished	22 April 2021 11:44 AM	22 April 2021 11:44 AM	42 secs	7.50	✗ 0.00	✓ 2.50	✓ 2.50	✓ 2.50
Finished	22 April 2021 11:45 AM	22 April 2021 11:45 AM	10 secs	10.00	✓ 2.50	✓ 2.50	✓ 2.50	✓ 2.50
Finished	22 April 2021 11:45 AM	22 April 2021 11:45 AM	30 secs	10.00	✓ 2.50	✓ 2.50	✓ 2.50	✓ 2.50
Finished	22 April 2021 7:12 PM	22 April 2021 7:15 PM	2 mins 52 secs	10.00	✓ 2.50	✓ 2.50	✓ 2.50	✓ 2.50
Finished	4 May 2021 4:15 AM	4 May 2021 4:23 AM	8 mins 20 secs	7.50	✗ 0.00	✓ 2.50	✓ 2.50	✓ 2.50
Finished	4 May 2021 4:24 AM	4 May 2021 4:24 AM	13 secs	10.00	✓ 2.50	✓ 2.50	✓ 2.50	✓ 2.50

Fig. 2. The interface of the Moodle teaching backend management system.

From early observations of the overall data, we found that the frequency of students logging into the Moodle system for learning was lowest at 5:00, so we used 5:00 as the starting point of a new day. For this study, we used a 28-hour time format, dividing the entire 24 hours into 5:00 to 28:00. The period from 24:00 to 28:00 is defined as the late-night time period. The advantage of switching from a 24-hour format to a 28-hour format is that the time variable responds more directly to changes in the learning time period. Table 1 shows a summary of all variables.

Table 1. Summary of variables

Variable	Description
V_i (i = 1, ..., 8)	The learning time period when the current lesson had been completed for Lessons 1-8
P	Study place (C: campus, H: home)
N_h	Number of homework submissions
N_e	Number of extra learning assignments submissions
T_v	Video learning content finished times.
G	School group
S	Score in the final examination
T_r	Review times
L	The study was completed between 24:00 and 28:00

4 Results

4.1 Data summary

As can be seen in Table 2, The median indicates that the period from afternoon to evening can represent the average study habits of students in online learning. Learning in the late-night period is not the dominant trend.

Table 2. Summary data of the learning time periods (5:00 to 28:00)

Variable	Maximum	Minimum	Mean	SD	Median	Variance
V_1	28:00	9:00	14.667	3.637	14:00	13.224
V_2	28:00	8:00	16.788	5.187	15:00	26.904
V_3	27:00	5:00	17.465	5.693	18:00	32.415
V_4	28:00	5:00	17.222	5.991	18:00	35.889
V_5	28:00	6:00	17.626	5.665	18:00	32.094
V_6	27:00	9:00	17.061	5.146	17:00	26.486
V_7	27:00	7:00	18.909	5.368	21:00	28.818
V_8	28:00	8:00	17.172	5.566	18:00	30.98

As can be seen in Fig. 3, the summary data for all learning time periods. The peak time periods for online learning courses are concentrated at 14:00 and 23:00. This represents the preferred time periods for students to study online.

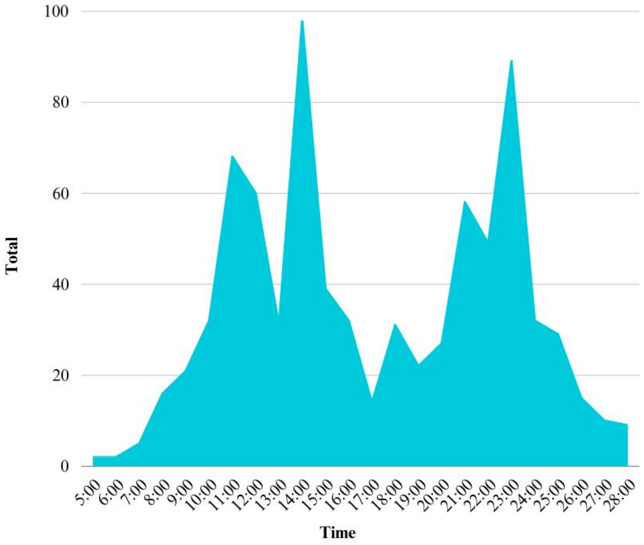


Fig. 3. Summary data for learning time periods.

4.2 Pearson Correlation Analysis

Pearson correlation coefficient analysis is an analytical method used to explore the degree of linear correlation between two variables. The results of the analysis can be obtained when the following conditions must be met: both variables were continuous and from the same sample; there was a linear relationship between the two variables; and the two variables were binary normally distributed or approximately normally distributed.

Table 3. Pearson Correlation analysis results

		T_v
	Coefficient	-0.024
L	*p*-value	0.813
	Sample size	99

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

The correlation analysis was used to analyze the correlation between L and T_v, and the Pearson correlation coefficient was used to indicate the strength of the correlation. As can be seen in Table 3, the value of the correlation coefficient is -0.024, which is approximately equal to 0, and the p-value is 0.813 > 0.05, indicating that there is no correlation between L and $T_{v.}$ The result of the analysis shows the students who preferred to study during the late night completed their study tasks as usual, and there was no impact on the learning motivation.

4.3 K-prototype cluster analysis

Cluster analysis explores and discovers the categorization of data and finds the characteristics of each category. When there are categorical data in the clustered categories, the K-prototype algorithm is used for cluster analysis. As can be seen in Table 4, the final clustering result divides the sample into 2 groups: Group A and B. The size of the samples in the two groups is 46.46% and 53.54% of the overall sample size respectively. The distribution of the two groups of people is comparatively equal, indicating that the clustering model is as good as it can be.

Table 4. Summary of K-prototype cluster results

Cluster Group	*n*	Percent（%）
A	46	46.46%
B	53	53.54%
Total	99	100%

The chi-squared test was used to find the differences between the qualitative data and was utilized to investigate whether there was a significant difference between the cluster groups for L. As can be seen from Table 5, Cluster Groups present a 0.01 level of significance for L ($\chi^2 = 8.592$, $p = 0.003 < 0.01$), indicating that both Group A and B have significant differences for all of V_{17}. The results of the chi-square test illustrate that there is a significant difference in the tendency of Group A and Group B to study late at night.

Table 5. Summary of chi-square analysis results

Items	Categories	Cluster groups (%)		Total	$\chi 2$	*p*
		1.0	2.0			
L	A	0 (0.00)	9 (16.98)	9 (9.09)		
	B	46 (100.00)	44 (83.02)	90 (90.91)	8.592	0.003**
Total		46	53	99		

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

The results of the chi-square analysis revealed between-group differences in the cluster groups. To further observe the differences between Group A and Group B, We plotted the differences between the two groups regarding learning time periods as shown in Fig. 4, The peak frequency of overall learning time periods for Group A can be seen in the morning and afternoon, while the peak frequency of overall learning time periods for Group B can be seen in the evening. It can be assumed that Group A prefers to study in the morning and afternoon, while Group B prefers to study in the evening.

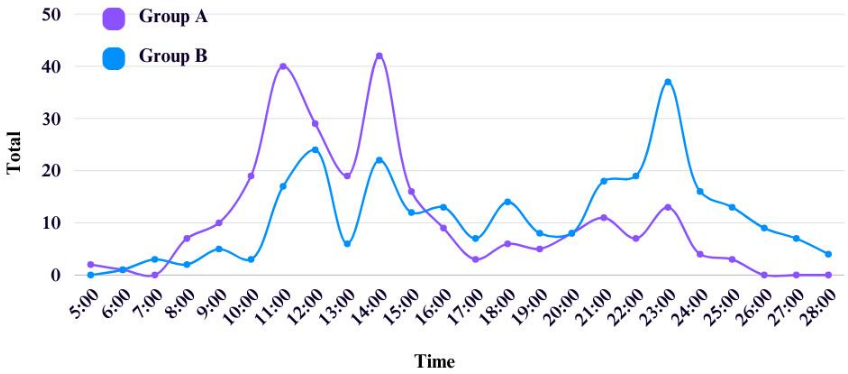


Fig. 4. Differences between groups in the learning time periods.

The ANOVA was used to investigate the differences in the cluster groups of *S*, as can be seen in Table 6, the different cluster groups of *S* showed a level of significance of 0.01 ($F = 12.853$, $p = 0.001$). Comparing the differences in the ANOVA analyzed results, it can be seen that the mean value of group A (85.43), will be significantly higher than the mean value of group B (72.45). This represents a significantly better academic performance of Group A than Group B.

Table 6. Summary of one-way ANOVA results

Items	Categories	n	Mean	SD	*F*	*p*
	Group A	46	85.43	9.59		
S	Group B	53	72.45	22.86	12.853	0.001**
	Total	99	78.48	19.03		

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

By specifically analyzing between-group differences in clustering classification results, we found that group A preferred to study in the morning and afternoon, and group B preferred to study at night. Group A performed better than Group B in terms of academic performance.

5 Conclusion

The results obtained in this study reveal the tendency of students in online education to study in time periods when they can flexibly arrange their learning time. The log data from Moodle showed that the peak learning time periods for all students were clustered around 14:00 and 23:00.

According to the results of Pearson's correlation analysis, the video learning content completion rate of students studying in the late-night time period showed no difference from that of students studying in other time periods, illustrating the fact that studying late at night had the same learning motivation compared to the daytime.

Cluster analysis categorized the overall sample into two groups, and the two groups of students differed in their preferences for learning time periods. According to the results of the ANOVA, the group of students who preferred to study during the daytime performed better in academic performance, suggesting the possibility that study habits preferring to study more during the daytime may have contributed to students' academic improvement.

The limitation of this study is that the Moodle system records relatively few types of data. Obtaining more types of data on students' online learning behaviors would enable a better understanding of students' time preferences in online learning. The findings of this study have important implications for further adapting online course design. In subsequent research endeavors, we will further investigate the significance of learning motivation on students' academic performance.

Acknowledgments

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

References

1. Yousef, A. M. F., Chatti, M. A., Schroeder, U., Wosnitza, M., & Jakobs, H.: MOOCs-A Review of the State-of-the-Art. In International Conference on Computer Supported Education **2**, 9-20 (2014)
2. Szymkowiak, A., Melović, B., Dabić, M., Jeganathan, K., & Kundi, G. S.: Information technology and Gen Z: The role of teachers, the internet, and technology in the education of young people. Technology in Society 65, 101565 (2021)
3. Mpungose, C. B.: Emergent transition from face-to-face to online learning in a South African University in the context of the Coronavirus pandemic. Humanities and social sciences communications **7**(1), 1-9 (2020)

4. Du, X., Zhang, M., Shelton, B. E., & Hung, J. L.: Learning anytime, anywhere: a spatio-temporal analysis for online learning. Interactive Learning Environments **30**(1), 34-48 (2022)
5. Castro, M. D. B., & Tumibay, G. M.: A literature review: efficacy of online learning courses for higher education institution using meta-analysis. Education and Information Technologies **26**(2), 1367-1385 (2021)
6. Vanichvatana, S., Tantanawat, S., Teawanit, T., & Siripat, A.: Characteristics of Midnight Libraries: Cases of Thailand National Research Universities. https://uruae.org/siteadmin/upload/UH0117408.pdf, last accessed 2023/10/25
7. Kronfeld-Schor, N., & Einat, H.: Circadian rhythms and depression: human psychopathology and animal models. Neuropharmacology **62**(1), 101-114 (2012)
8. Bustamante, C. M. V., Coombs III, G., Rahimi-Eichi, H., Mair, P., Onnela, J. P., Baker, J. T., & Buckner, R. L.: Precision Assessment of Real-World Associations Between Stress and Sleep Duration Using Actigraphy Data Collected Continuously for an Academic Year: Individual-Level Modeling Study. JMIR Formative Research **8**(1), e53441 (2024)
9. Master L, Nahmod NG, Mathew GM, Hale L, Chang AM, Buxton OM.: Why so slangry (sleepy and angry)? Shorter sleep duration and lower sleep efficiency predict worse next-day mood in adolescents. Journal of Adolescence **95**(6), 1140-1151 (2023)
10. Fallone, G., Owens, J. A., & Deane, J.: Sleepiness in children and adolescents: clinical implications. Sleep medicine reviews **6**(4), 287-306 (2002)
11. Dewald, J. F., Meijer, A. M., Oort, F. J., Kerkhof, G. A., & Bögels, S. M.: The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: A meta-analytic review. Sleep medicine reviews **14**(3), 179-189 (2010)
12. Curcio, G., Ferrara, M., & De Gennaro, L.: Sleep loss, learning capacity and academic performance. Sleep medicine reviews **10**(5), 323-337 (2006)
13. Bartel, K. A., Gradisar, M., & Williamson, P.: Protective and risk factors for adolescent sleep: a meta-analytic review. Sleep medicine reviews **21**, 72-85(2015)
14. Romero, C., & Ventura, S.: Educational data mining: A survey from 1995 to 2005. Expert systems with applications **33**(1), 135-146 (2007)
15. Aulakh, K., Roul, R. K., & Kaushal, M.: E-learning enhancement through educational data mining with Covid-19 outbreak period in backdrop: A review. International journal of educational development **101**, 102814 (2023)
16. Abdelkader, H. E., Gad, A. G., Abohany, A. A., & Sorour, S. E.: An efficient data mining technique for assessing satisfaction level with online learning for higher education students during the COVID-19. IEEE Access **10**, 6286-6303 (2022).
17. Jalota, C., & Agrawal, R.: Analysis of educational data mining using classification. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 243-247 IEEE (2019)
18. Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M.: Clustering for improving educational process mining. In Proceedings of the fourth international conference on learning analytics and knowledge, 11-15 (2014)