Modelling Student Performance in E-Learning STEM Education using Logistic Regression Approach

Nurulhuda Ramli

Abstract – **Predicting student performance in online STEM education remains challenging due to the heavy reliance in existing literature on static, pre-course learner characteristics. This study addressed this gap by developing a predictive model that integrates both individual e-learning characteristics and in-process e-learning behaviours. Utilizing the Open University Learning Analytics Dataset (OULAD), Logistic Regression (LR) models were developed to quantify the impact of these variables. The analysis confirmed the high influence of active e-learning engagement in predicting student outcomes. The student's score on online assessment was identified as the universal and most robust predictor across all achievement categories (Distinction, Pass, Fail), highlighting its central role in conceptual mastery. Crucially, the e-learning engagement level was found to be highly significant for achieving top-tier success. While the model achieving a moderate overall accuracy of 53%, category-specific metrics revealed significant performance variability. The actionable insights derived from this study can inform targeted pedagogical interventions and refine the understanding of complex e-learning dynamics in STEM environments.**

Keywords – **Predictive analysis, E-learning behaviour, STEM course, Logistic Regression, Data Mining**

I. INTRODUCTION

STEM education — an acronym for Science, Technology, Engineering, and Mathematics has received widespread international attention as the crucial mechanism for responding to the challenges of careers in the intelligent age and nurturing innovative talent. This renewed global focus on integrated learning aligns strongly with the principles of Education 4.0, which advocates for flexible, personalized, and technology-enhanced instruction, and directly supports the achievement of the United Nations Sustainable Development Goals (SDGs). In the higher education setting, the implementation of STEM has gained critical prominence as the primary vehicle for cultivating advanced competencies and ensuring graduates are truly 'job-ready' for the dynamic, highly technical marketplace (Johnson et al., 2021). Higher education institutions are increasingly focused on leveraging integrated STEM courses to equip students with the sophisticated critical thinking, complex problem-solving skills, and adaptability required of future talent and industry leaders (Alali, 2024).

The constant evolution of digital technology, in parallel, presents significant opportunities to substantially enhance the quality and accessibility of STEM education.

The constant evolution of digital technology, in parallel, presents significant opportunities to substantially enhance the quality and accessibility of STEM education, thereby promoting greater global equity. Specifically, the rise of Massive Open Online Courses (MOOCs) and other dedicated online platforms has fundamentally reshaped traditional learning models. These virtual environments transcend conventional time and space constraints, offering learners personalized pacing, sophisticated mechanisms for recording learning processes (analytics), and immediate access to rich, often curated, digital resources (Siemens, 2018). Consequently, the development and application of robust measurement and prediction mechanisms for assessing student performance in online STEM courses has become a major focus area among educational technology researchers (Baker & Inventado, 2023).

Nurulhuda Ramli, School of Distance Education, Universiti Sains Malaysia, Malaysia (Email address: rnurulhuda@usm.my).

II. PROBLEM STATEMENT

Even though many studies have been published in the field of learning achievement analysis, demonstrating positive results in improving overall learner performance and optimizing course construction, significant limitations and research gaps persist. Firstly, most of the STEM education improvement literature focuses heavily on building and validating curricula within traditional, face-to-face (offline) settings. This resulting in a scarcity of research in understanding and enhancing the learning performance of students in online STEM courses (Su et al., 2022; Zhang et al., 2023). Secondly, when existing literature does attempt to predict learners' online performance of STEM education, many existing predictive models rely predominantly on individual e-learning characteristics (e.g., prior knowledge, demographics), with limited consideration given to dynamic in-process online engagement behaviours that occur during the e-learning process (Brahim, 2022; Wang & Yu, 2025).

To this end, the study seeks to answer the following questions:

1. What are the essential individual e-learning characteristics and e-learning behaviours that exert a statistically significant influence on the prediction of learner performance outcomes in STEM courses?
2. To what extent can the collected in-process e-learning behaviour data effectively predict final performance in STEM courses?

Addressing these questions is essential for informing data-driven interventions that will enhance student e-learning performance and refine pedagogical strategies in STEM education.

III. LITERATURE REVIEW

The research on predicting student performance typically revolves around two primary aspects: the

attributes (variables) used as input and the prediction methods (algorithms) employed (Shairi et al., 2015). Historically, the most utilized single indicator for predicting overall student success at the university level has been the Cumulative Grade Point Average (CGPA) (Shairi et al., 2015), a factor employed widely in numerous studies (e.g., (Christian & Ayub, 2014; Nguyen & Haddawy 2007). Beyond CGPA, researchers commonly use static, course-specific criteria such as quiz grades, assessments, lab work, and final exam marks to predict achievement (e.g., (Arsad et al., 2013)). Further static input variables, such as basic student demographics, are also common. Examples of these individual e-learning characteristics (or propensity indicators) include socioeconomic status (Romero et al., 2016), historical academic records (Nawang et al., 2017), and gender (Keogh, E. J. & Mueen, 2017). Although the predictive models established using these static attributes often achieve reasonable performance, they generally overlook the crucial role of active, in-course learning behaviour records. In contrast, e-learning behaviour data can accurately capture the time and energy students spend on a specific course. These dynamic, in-process indicators—such as the frequency of access to course materials (Marquez-Vera et al., 2016), duration of video interaction, and the frequency of online discussions (Marbouti et al., 2016), student engagement using log data (Kizilcec et al., 2013; Milligan et al., 2013) —provide a granular view of the actual learning process. Because these behavioural records offer greater insight into student effort and engagement, some studies have successfully combined both individual e-learning characteristics and in-process e-learning behaviour indicators to enhance learning prediction accuracy (Zhao et al., 2021).

In e-learning performance prediction research, the selection of predictive indicators is critically important. Researchers commonly employ data mining techniques, primarily classification and regression, to build models for student performance prediction (Shairi et al., 2015). Classification techniques are used when the outcome variables are categorical or discrete (e.g., Pass/Fail, Drop/Complete), whereas regression techniques are used for numerical or continuous outcome variables (e.g., final course score). In the context of higher education, classification is the most widely used data mining approach for prediction (Aldowah et al., 2019).

A diverse array of machine learning classification algorithms have been successfully applied in this field. Traditional algorithms, such as K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Trees (DT), and Logistic Regression (LR) are frequently used to predict student performance. For instance, Asif et al. (2017) utilized the DT algorithm to predict student performance across a four-year curriculum, while Mishra and Kumar (2014) employed various implementations of the DT technique to model performance focused on students' social integration. Furthermore, Quadri and Kalyankar (2010) utilized DT specifically to predict student dropout rates. Other researchers have explored different approaches: Ahuja and Kankane (2017) used the KNN algorithm to predict academic results based on previous academic performance and non-academic factors. Similarly, Aziz et al. (2015) selected five static parameters (race, gender, family income, college enrollment mode, and average grade point) and applied the NB classifier. Jiang et al. (2014) built a robust predictor based on LR, combining early homework performance and social interaction behaviours. These varied applications demonstrate the broad scope of predictive modeling used to understand and forecast student success.

IV. METHOD

Introduction to the Dataset

The Open University Learning Analytics dataset (OULAD), which includes a portion of the Open University (OU) student data from 2013 and 2014, was utilized as the primary data source for this study. The dataset is student-oriented, encompassing information on student module registrations, demographics, and the outcomes of student assessments for each student-module-presentation triplet. Crucially, a log of each student's daily activity and interactions within the Virtual Learning Environment (VLE) is also included. The complete OULAD dataset covers 32,593 registered students and 22 module presentations, and is freely accessible (e.g., at https://analyse.kmi.open.ac.uk/open_dataset). The dataset holds accreditation from the Open Data Institute (http://theodi.org/) (Kuzilek et al., 2017).

TABLE I: THE VARIABLES USED IN ANALYSIS

Variables	Description and type
Gender	Student's gender. (Binary)
Age	A band of student's age. (Nominal)
Disability	Whether the student has announced a disability is indicated. (Binary)
Highest education	The greatest degree of education a student has at the time of the module presentation. (Nominal)
Studied credits	The total credits earned for the modules the student is enrolled in. (Nominal)
IMD band	Band of statistical data (Nominal)
Assessment score	The result of assessment for the student. (Nominal)
VLE engagement level	The interaction of student based on number of times the student interacted with the material. (Nominal)
Final result	The final result of the student for the module presentation. (Nominal)

For this present study, the analysis was focused on a subset of 2,564 learners registered for a single STEM course, coded as EEE. The input variables selected for the study fall into two categories: individual e-learning characteristics (gender, age, disability, highest education, studied credit, IMD band, and assessment score) and in-process e-learning behaviour (VLE engagement level). The outcome, or output value, is the categorical final result (Pass, Distinction, Fail, and Withdrawn). List of variables used is presented in Table 1. Before applying classification techniques, the raw dataset underwent comprehensive pre-processing and transformation steps to convert non-numerical data into appropriate numerical variables and

classify them into categories suitable for predictive analysis.

Logistic Regression (LR)

LR analysis is used to find the best model to describe the relationship between the dependent variable and the independent variable. Called logistic regression, because in this regression analysis the formation of the model is based on logistic curves. The resulting value of the logistic regression equation is the chance of the event being used as a measure for classification (Hosmer & Lemeshow, 2000). By fitting data to a logistic function, the type of regression known as "logistic regression" can predict the probability that an event will occur (Marrakchi et al., 2023). Similar to other types of regression analysis, logistic regression uses a number of predictor variables that may be categorical or numerical (Dreiseitl & Ohno-Machado, 2002). The maximum likelihood function for estimating model parameter is

$$I(\beta) = \prod \pi_0 (x_i)^{y0i} \pi_1 (x_i)^{y1i} \pi_2 (x_i)^{y2i} \text{ , with } \sum ij = 1$$

then the log function is a likelihood

$$I(\beta) = \sum y_{1ig1} (x_i) + y_{2ig2} (x_i) - \ln (1 + e^{g_1(x_i)} + e^{g_2(x_i)}$$

Test the significance of parameters simultaneously with the test statistic G or likelihood ratio, $G = 12\ln [L_0/L_a]$ where L_0 is Likelihood without independent variables, and a is Likelihood with independent variables. Meanwhile the partial test using a Wald test, $W = (\hat{\beta}_i / SE(\hat{\beta}_i))^2$.

V. FINDINGS

Strength of relationships

In this study, the performance of an online STEM course was predicted based on collected e-learning behaviour and student characteristics. Prior to predictive modeling, a Cramér's V test and the chi-square test of independence were utilized to describe the strength and significance of the relationship between the dependent variable (Final Result) and all independent categorical variables. The results of this preliminary analysis will inform the subsequent predictive modelling. The findings of this analysis are summarized in Table II.

TABLE II: ASSOCIATION ANALYSIS AND CHI-SQUARE TEST FOR INDEPENDENT VARIABLES VS FINAL RESULT

Variables	Cramérs V	χ^2	Sig.
Gender	0.0356	3.2529	0.3542
Age	0.0364	6.8056	0.3392
Disability	0.0418	4.4754	0.2145
Highest education	0.0806	49.9711	1.4135e-06*
Studied credits	0.1194	109.6936	5.3534e-14*
IMD band	0.0943	58.8899	9.1478e-05*
Assessment score	0.2919	615.1786	1.1557e-13*
VLE engagement level	0.3092	735.1435	1.31545e-15*

Note: *Association significant at the 0.05 level

The results reveal that several key student characteristics (Highest Education, Studied Credits, IMD Band and Assessment score) and e-learning behavioural indicators (VLE Engagement Level) have a statistically significant relationship with the final course result (Sig. < 0.05). VLE Engagement Level (Cramér's V = 0.3092) demonstrate the strongest associations with the final result. This suggests that e-learning behavioural factor highly influential in predicting outcome. Due to the lack of significance for some independent variables, these variables were excluded in the subsequent LR analysis.

Prediction analysis using LR

The LR analysis for predicting student performance was conducted in two phases: the simultaneous test and the partial test. The simultaneous test (or omnibus test of model coefficients) is a significance test used to determine whether the inclusion of a specific variable (or a block of variables) significantly improves the overall fit of the model compared to a model without that variable, which is also known as baseline model. This test utilizes the G or likelihood ratio test. The baseline model (also known as the null model or the intercept-only model) is a statistical model that includes only the intercept term (a constant value) and none of the independent predictor variables. The results for the key predictor variables are summarized in Table III.

TABLE III: SIMULTANEOUS LR TEST FOR PREDICTOR VARIABLES ON FINAL RESULT

Effect	$\widehat{\beta}_l$	G (Likelihood ratio statistic)	Sig.
Intercept	0.081	5901.21	
Highest education	-0.171	5883.53	0.00058*
Studied credit	0.461	5909.56	0.0000*
IMD band	-0.098	5885.85	0.00019*
Assessment score	0.892	6178.55	0.00000*
VLE engagement level	0.0086	6087.39	0.00000*

Note: * Significant at the 0.05 level

The results in Table III demonstrate the significant contribution of each variable to the predictive power of the LR model. The high values of the G and the corresponding extremely small significance values (Sig. < 0.05) for all listed variables indicate that every variable tested significantly improves the model's ability to predict the final course result when compared to a baseline model.

The partial test (using the Wald statistic) examines the unique, individual contribution of each predictor variable to the model after accounting for the influence of all other variables. For this partial test, the reference category was set to "Withdrawn". The significance value (Sig.) for the Wald statistic indicates whether a predictor makes a statistically significant contribution to the prediction of a specific outcome category. The results are

summarized in Table IV, which reveals that the individual predictive power of variables is highly dependent on the target outcome category.

The Assessment Score emerges as the most consistent and universally significant predictor across all three achievement levels (Distinction, Pass, and Fail, all Sig. < 0.05), confirming its central role in determining the student's final result. For the final result of Distinction, the model demonstrates that all predictors are statistically significant,

TABLE IV: PARTIAL LR TEST FOR PREDICTOR VARIABLES ON FINAL RESULT

Final result	Effect	$\widehat{\beta}_i$	Wald statistic	Sig.
Distincti-on	Intercept	-1.143	24.59	7.1e-1
	Highest education	0.081	1.003	0.3165
	Studied credit	-0.171	4.261	0.039*
	IMD band	-0.098	13.009	0.000*
	Assessment score	1.039	213.27	2.65e-5*
	VLE engagement level	0.461	64.17	1.14e-2*
Fail	Intercept	0.819	19.54	9.86e-1
	Highest education	-0.048	0.47	0.493
	Studied credit	-0.087	1.568	0.211
	IMD band	-0.042	3.26	0.0711*
	Assessment score	0.892	198.03	5.63e-5*
	VLE engagement level	0.0086	0.032	0.858
Pass	Intercept	-1.058	24.32	8.17e-8
	Highest education	0.168	5.039	0.025*
	Studied credit	0.183	6.5	0.0108*
	IMD band	-0.09	12.372	0.00044*
	Assessment score	0.815	147.8	5.26e-3*
	VLE engagement level	0.46	73.808	8.61e-2*

Note: * Significant at the 0.05 level

except Highest Education (Sig. 0.3165), suggesting that achieving the highest grade requires a favorable combination of prior standing, continuous assessment performance, and high VLE engagement. Conversely, predicting Fail relies almost exclusively on the Assessment Score and the IMD band (Sig < 0.05), implying that once assessment performance drops, behavioural efforts are insufficient to mitigate the risk of failure. Finally, the Pass outcome is significantly influenced by all student characteristics and e-learning behaviour variables, suggesting that this mid-range success category is the most robustly predicted by the full set of factors.

LR performance analysis

Table 5 summarizes the performance of the LR classification model using standard metrics: Precision, Recall, F1-Score, and overall Accuracy. The model achieved an overall accuracy of 53%. However, analysis of the category-specific metrics reveals significant performance variability. The model's strongest performance is in predicting the 'Pass' outcome, achieving a high Recall of 83% and the best F1-Score (0.68), indicating effectiveness in correctly identifying successful students. Conversely, the model shows significant weakness in classifying the high-risk categories. The 'Fail' category has low Recall (25%) and Precision (39%) (with F1-Score 0.30), meaning the model missed many actual failing students.

TABLE V: PERFORMANCE ANALYSIS

Output category	Precision	Recall	F1-Score	Accuracy
Overall				0.53
Distinction	0.47	0.37	0.41	
Fail	0.39	0.25	0.30	
Pass	0.57	0.83	0.68	
Withdraw	0.45	0.17	0.25	

The 'Withdraw' category shows the lowest performance, with a Recall of only 17%, making it poor at identifying potential withdrawals. Prediction for 'Distinction' is moderate (F1-Score 0.41). In summary, the LR model is best suited for identifying students who will achieve a mid-range 'Pass' result, but it lacks the necessary sensitivity and precision to reliably predict the crucial high-risk categories of 'Fail' and 'Withdraw.'

VI. DISCUSSION

The primary objective of this study was to address the critical gaps in online STEM education research by examining the predictive power of individual e-learning characteristics and in-process e-learning behaviours on final course outcomes. The results strongly validate the utility of incorporating granular behavioural data alongside traditional student attributes for predicting student performance.

In response to the first research question, the association analysis confirmed that in-process e-learning behaviours, demonstrate the strongest statistical relationship with the final result. This finding supports the argument that mere reliance on non-significant individual characteristics is insufficient for meaningful prediction in modern e-learning environments. The high influence of e-learning engagement, and ongoing assessment performance are more indicative of student success in online STEM courses than historical background factors such as prior academic achievement or demographics (Kim et al., 2019; Rajabalee & Santally, 2020; Soffer & Cohen, 2019).

Addressing the second research question, the LR analysis provided deep insights into how these variables predict specific outcome categories. The simultaneous test confirmed that including both individual e-learning characteristics and in-process engagement behaviours significantly improves prediction of final course outcomes in online STEM education, as supported by Hussain et al. (2018) and Al-Azawei et al. (2020). Meanwhile, continuous assessment results consistently emerge as the most robust and universal indicator of student success across outcome categories. This result confirms the critical importance of continuous, high-stakes evaluation as a central driver of the final result (Jawad et al., 2022). Crucially, the in-process e-learning behaviour (VLE Engagement Level) was found to be a significant predictor for Distinction. As Goode et al. (2022) and Kobicheva

(2022) suggested that high behavioural engagement is an effort multiplier necessary for achieving top performance. Conversely, this same behavioural measure lacked significant predictive power in the 'Fail' category, underscoring the need for intervention efforts to focus on diagnosing and resolving conceptual difficulties, rather than just relying on participation nudges.

VII. CONCLUSION (OR LIMITATION OR SUGGESTION FOR FURTHER STUDIES)

This study successfully addressed the critical research gaps in online STEM education by developing a predictive model that integrated both individual e-learning characteristics and in-process e-learning behaviours. Utilizing the OULAD dataset, the analyses confirmed the high influence of e-learning engagement in predicting student performance. While certain static factors like Studied Credit and IMD Band maintain significance, the Assessment Score serves as the universal and most robust predictor across all achievement categories, highlighting its central role in measuring conceptual mastery. Crucially, the VLE Engagement Level — the primary in-process e-learning behaviour examined was found to be highly significant for achieving Distinction, functioning as an effort multiplier necessary for top-tier success. While the model provides valuable, actionable intelligence, achieving an overall accuracy of 53%, its low sensitivity in identifying high-risk students, particularly those who 'Fail' or 'Withdraw,' presents a significant limitation. Therefore, future research must prioritize improving predictive accuracy by focusing on exploring more sophisticated, potentially non-linear classification algorithms to enhance the model's sensitivity and precision for these critical high-risk categories. Furthermore, future studies should incorporate a wider array of in-process e-learning behaviours (such as the duration of video viewing, pattern shifts in weekly forum participation, and peer interaction quality) to more effectively capture the complex e-learning dynamics specific to STEM online environments.

ACKNOWLEDGEMENT

This work was supported by Universiti Sains Malaysia under R502 - KR - ARU005 - 0000000588 - K134 GGPM-2022-040 (Geran Penyelidikan Akademik 2024).

REFERENCES

Ahuja, R., & Kankane, Y. (2017). Predicting the probability of student's degree completion by using different data mining techniques. *2017 Fourth International Conference on Image Information Processing (ICIIP)* (pp. 474–477). IEEE. https://doi.org/10.1109/ICIIP.2017.8313763

Alali, R. A. (2024). Enhancing 21st century skills through integrated STEM education using project-oriented problem-based learning. *GeoJournal of Tourism and Geosites*, 53(2), 528–534. https://doi.org/10.30892/gtg.53205-1217

Al-Azawei, A., & Al-Masoudy, M. (2020). Predicting Learners' Performance in Virtual Learning Environment (VLE) based on Demographic, Behavioural and Engagement Antecedents. *International Journal of Emerging Technologies in Learning*, 15(9), 60–75. https://doi.org/10.3991/ijet.v15i09.12691

Aldowah, H., Al-Samarraie, H., & Fauzy, M. W. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, 37, 13–49.

Arsad, M. P., Buniyamin, N., & Manan, A. J. (2013). A neural network students' performance prediction model (NNSPPM). *2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)* (pp. 1–5). IEEE.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 113, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Baker, R. S. J. d., & Inventado, P. S. (2020). Educational data mining and learning analytics. In A. T. Corbett, W. K. C. M. L. W. T. H. E. (Eds.), *The Cambridge Handbook of Computing Education Research* (pp. 524–564). Cambridge University Press.

Brahim, G. (2022). Predicting student performance from online engagement activities using novel statistical features. *Arabian Journal for Science and Engineering*, 47(11), 10225–10243. https://doi.org/10.1007/s13369-021-06548-w

Christian, M. T., & Ayub, M. (2014). Exploration of classification using NBTree for predicting students' performance. *2014 International Conference on Data and Software Engineering (ICODSE)* (pp. 1–6). IEEE.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, 35(5-6), 352–359.

Goode, E., Nieuwoudt, J., & Roche, T. (2022). Does online engagement matter? The impact of interactive learning modules and synchronous class attendance on student achievement in an immersive delivery model. *Australasian Journal of Educational Technology*, 38(6), 1–19. https://doi.org/10.14742/ajet.7929

Hosmer, D. W., & Lemeshow, S. (2000). Applied logistic regression (2nd ed.). John Wiley & Sons Inc.

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. (2018). Student engagement predictions in an e-Learning system and their impact on student course assessment scores. *Computational Intelligence and Neuroscience*, 2018, 1–12. https://doi.org/10.1155/2018/6347186

Jawad, K., Shah, M., & Tahir, M. (2022). Students' academic performance and engagement prediction in a virtual learning environment using Random Forest with data balancing. *Sustainability*, 14(22), 1–18. https://doi.org/10.3390/su142214795

Jiang, S., Williams, A. E., Schenke, K., Warschauer, M., & O'Dowd, D. (2014). Predicting MOOC performance with week 1 behaviour. *Proceedings of the 7th International Conference on Educational Data Mining, EDM 2014* (pp. 273–275). International Educational Data Mining Society (IEDMS).

Johnson, L., Becker, S. A., Cummins, M., Estrada, V., Freeman, A., & Hall, C. (2016). *NMC horizon report: 2016 higher education edition*. The New Media Consortium.

Keogh, E. J., & Mueen, A. (2017). Curse of dimensionality. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 314–315). Springer. https://doi.org/10.1007/978-1-4899-7687-1_192

Kim, H., Hong, A., & Song, H. (2019). The roles of academic engagement and digital readiness in students' achievements in university e-learning environments. *International Journal of Educational Technology in Higher Education*, 16(1), 1–18. https://doi.org/10.1186/s41239-019-0152-3

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. *Proceedings of 3rd International Conference on Learning Analytics and Knowledge* (pp. 170–179). ACM.

Kobicheva, A. (2022). Comparative study on students' engagement and academic outcomes in live online learning at university. *Education Sciences*, 12(6), 1–16. https://doi.org/10.3390/educsci12060371

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics dataset. *Scientific Data*, 4(1), 1–8. https://doi.org/10.1038/sdata.2017.171

Marbouti, F., Diefes-Dux, H., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, 103, 1–15. https://doi.org/10.1016/j.compedu.2016.09.005

Marquez-Vera, C., Prior, D., De La Calle, D. J., & Fernández-Balch, G. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, 33(1), 107–124. https://doi.org/10.1111/exsy.12135

Marrakchi, N., Bergam, A., Fakhouri, H., & Kenza, K. (2023). A hybrid model for predicting air quality combining Holt–Winters and deep learning approaches: A novel method to identify ozone concentration peaks. *Mathematical Modeling and Computing*, 10(4), 1154–1163.

Milligan, C., Littlejohn, A., & Margaryan, A. (2013). Patterns of engagement in connectivist MOOCs. *Journal of Online Learning and Teaching*, 9(2), 149–159.

Mishra, T., Kumar, D., & Gupta, S. (2014). Mining students' data for prediction performance. *2014 Fourth International Conference on Advanced Computing & Communication Technologies* (pp. 255–262). IEEE.

Nawang, H., Makhtar, M., & Shamsudin, S. (2017). Classification model and analysis on students' performance. *Jurnal Fundamental dan Aplikasi Sains*, 9(6S), 869–885. https://doi.org/10.4314/jfas.v9i6s.65

Nguyen Thi Ngoc Hien, & Haddawy, P. (2007). A decision support system for evaluating international student applications. *2007 37th Annual Frontiers In Education Conference—Global Engineering: Knowledge Without Borders, Opportunities Without Passports* (pp. F2A-1–F2A-6). IEEE.

Quadri, N. M. M. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*, 10(2), 2–5.

Rajabalee, Y., & Santally, M. (2020). Learner satisfaction, engagement and performances in an online module: Implications for institutional e-learning policy. *Education and Information Technologies*, 26(3), 2623–2656. https://doi.org/10.1007/s10639-020-10375-1

Romero, C., Cerezo, R., Bogarín, A., & Sànchez-Santillán, M. (2016). Educational process mining: A tutorial and case study using moodle data sets. *Data Mining and Learning Analytics Applications in Educational Research*, 2, 1–28.

Shahiri, M. A., Husain, W., & Rashid, A. N. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414–422.

Siemens, G. (2017). Learning analytics: The emergence of a discipline. *American Behavioural Scientist*, 61(10), 1205–1218.

Soffer, T., & Cohen, A. (2019). Students' engagement characteristics predict success and completion of online courses. *Journal of Computer Assisted Learning*, 35(3), 378–389. https://doi.org/10.1111/jcal.12340

Su, Y., Chang, C., Wang, C., & Lai, C. (2022). A study of students' learning perceptions and behaviours in remote STEM programming education. *Frontiers in Psychology*, 13, 962984. https://doi.org/10.3389/fpsyg.2022.962984

Wang, J., & Yu, Y. (2025). Machine learning approach to student performance prediction of online learning. *PLOS ONE*, 20(4), e0299018. https://doi.org/10.1371/journal.pone.0299018 (Note: Volume and issue numbers are based on the DOI and typical PLOS ONE structure for early 2025.)

Zhang, J., Qiu, F., Wu, W., Wang, J., Li, R., Guan, M., & Huang, J. (2023). E-Learning behaviour categories and influencing factors of STEM courses: A case study of the Open University Learning Analysis Dataset (OULAD). *Sustainability*, 15(10), 8235. https://doi.org/10.3390/su15108235

Zhao, L., Huang, S., Chen, Y., Yu, D., Wang, Y., Zhang, X., & Xu, D. (2021). Academic performance prediction based on multisource, multifeature behavioural data. *IEEE Access*, 9, 5453–5465. https://doi.org/10.1109/ACCESS.2020.3002791