Proceedings of the 4th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/83/2025.19951

Analyzing High School Students' Online Learning Behavior Through Data Mining: Personalized Pathway Construction and Outcome Prediction

Hanyun Zhang[1,a,*]

[1]*The Chinese University of Hong Kong, Hong Kong, China*
a. Zhanghanyun@szys.net
**corresponding author*

Abstract: The objective of this research is to use data mining and machine learning methods to track the behavior of high school students through online education platforms, so that we can learn trends, create personalized pathways, and forecast academic performance. We took engagement metrics including sessions, videos watched, and assignment completion rates from 1,000 high school students. We used a k-means clustering algorithm to break students into high, moderate and low engagement groups with different learning habits and academic performance patterns. Customised learning modules were then crafted for each cohort, tailoring content to maximise participation and facilitate learning. They also employed a decision tree model to predict performance which was 82% accurate with session length and assignment completion rates being among the best predictors. The results show that targeted interventions, rooted in engagement-based data, have the potential to positively impact learning experiences on digital platforms for high school students.

Keywords: Data mining, learning analytics, high school students, online learning, personalized learning pathways

1. Introduction

As technology rapidly evolved, it changed education, with the emergence of internet based education institutions that provide students around the globe with accessible and personalized learning opportunities. This has particularly affected high school students, who spend a great deal of time on digital channels that offer organized learning and interactive material outside the classroom. But online courses aren't the only way for students to take classes when they feel like it, and they also come with their own unique challenges, such as lack of physical interaction with peers and teachers. For high school students, who are typically at the start of developing self-directed learning abilities, knowing how to learn on the internet is vital to their academic achievement. Data mining and learning analytics are becoming essential for teaching as it allows students and teachers to mine enormous datasets from their interaction with these platforms. Data mining uses clustering, classification, and predictive modeling to create a methodical method for identifying patterns in student behaviour that were previously hard to spot. Such insights can guide personalised learning interventions, which would enable education platforms to tailor materials for learners' specific learning requirements and interests. Using clustering, for example, students can be clustered in terms of engagement metrics

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/83/2025.19951

(such as time spent on tasks, homework completion rates, and frequent interaction), which reveals unique behavioral profiles to facilitate personalized support [1].

While data-driven approaches are increasingly being used in higher education, there is little research specifically targeting the online learning behaviour of high-school students. Students in high school demonstrate diverse digital literacy and engagement abilities depending on age, interest, and learning style. This research is an attempt to fill that void by examining the online learning of high school students, putting them into engagement clusters, and applying predictive modelling to identify students at risk and improve the educational programs. In creating personalised learning experiences with student engagement data, this study aims to support the digital experience of high-school students in order to provide an environment for adaptive, data-driven education. In doing so, we intend to help foster a more personalized and supportive online learning community, adapted to the many needs of high-school students.

2. Literature Review

2.1. Data Mining in Education

Data mining revolutionises education because it allows educators and researchers to analyze massive sets of educational data and derive insights that aren't easily observable. It also employs techniques such as clustering, classification, predictive modelling to parse and model data in a way that old ways of doing so cannot. Using clustering, for example, you can group students into classes according to similar traits like learning intensity, learning style, or achievement score. Classification, in contrast, allows students to be organized by pre-established rules and is used to design adaptive programs for the various learners. The predictive analytics goes a step further by modelling the future, allowing educators to predict what students will learn, when they will likely drop out, and whether intervention is needed. Such strategies help schools, as well as the information they give, adopt data-driven strategies [2]. Data mining, then, provides a systematic way for educators to improve teaching and learning through an understanding of learners' differences and evidence-based adjustments to teaching and resource use.

2.2. Behavioral Patterns in Online Learning

High-school students show a range of behavioural trends when engaging with online learning experiences, and these differences impact their academic achievements and learning outcomes. These practices — including measures of time spent on different types of content, the order of topics covered, and how often and when they attempt tests — provide an insight into the student learning experience online. For example, students who take content seriously for longer periods of time can be viewed as more disciplined and committed, and may be associated with better grades. Likewise, sequence of topic-focused activity might reflect a student's conceptualisation of learning goals, and students who engage in a linear, progressive approach to learning are likely to retain more content than those engaging in random sequences [3]. Having a repeat score can show a student's commitment to learning, but a bit of occasional contact can mean time-management or understanding difficulties. It's important to have a grasp of these behaviour patterns, because they help to determine what kind of individuals in the student population have different learning preferences and needs. Through analysis of these trends, this research can provide data to customize learning experiences to adapt online learning to the behavioural diversity of high school students.

DOI: 10.54254/2753-8818/83/2025.19951

2.3. Personalized Learning Pathways

Personalised learning journeys are a developing education method that caters to the needs, capacities and learning styles of specific students. By arranging content and pace according to each individual student's progress, customized tracks make learning more dynamic and student-focused. Such arrows are possible thanks to the use of behavioural data, which tell us how students engage with content and what factors affect their learning and achievement. For example, high-flying students can be given extra accelerated material that makes the subject more challenging, and struggling students could get additional scaffolding and review materials. Online personalized using algorithms enables adaptive feedback at the moment of use so that each student is receiving what they require based on their behaviours [4]. This customisation both holds students' attention and helps them feel like they are in control of their learning, because information becomes more personalized for them. By creating these avenues, learning platforms can evolve from one-off collections of content to one-way environments that are adaptive to the unique demands of each student, improving learning and helping the diverse high school student body in a centralized, focused manner.

3. Methodology

3.1. Data Collection

Data were collected through one of the largest online education platforms catering to high school students, and from that data, a dataset of 1,000 students across multiple subjects was produced. These data points encompassed a range of behavioural measures, such as the number of sessions, the number of interactions with instructional videos, assignment completion, and quiz scores. Each of these measures were a way of measuring student participation, which could potentially reveal patterns of learning. For example, time spent in class offered a sense of commitment, and watching videos represented active learning. Graduation rates on assignments were carefully watched for consistency, and quiz grades were used as a proxy for understanding and retention. To stay within ethical guidelines, data were anonymous to avoid invading student privacy, and no identity was ever identifiable during analysis [5]. Furthermore, to make the data more robust, the data was preprocessed. Data normalization standardized the values of all the metrics to a standard scale, which prevented larger ranges of features from having an undue influence on analysis results. Outliers, detected by statistical techniques, were filtered out to eliminate any biases that could skew the clustering and predictive modeling. The clean data gave us a solid baseline for further analyses.

3.2. Clustering Analysis

A k-means clustering algorithm was used to efficiently classify student behaviors and build learning profiles. K-means clustering is an iterative algorithm to divide a dataset into k clusters, each of which has its own centroids – the average values of data points in the cluster. In this study, we selected three clusters (k=3) to demarcate the high, medium, and low engagement behaviors among students. The algorithm randomly placed all the students into clusters, then adjusted the tasks incrementally based on distance from cluster centroids, thus reducing within-cluster variation. Three clusters that emerged exhibited their own engagement patterns. 30 % of the sample belonged to cluster 1, or high engagement. They had extended sessions, participated frequently in the material, and did most of the homework regularly, suggesting a proactive learning style. Moderate engagement: Cluster 2 contained moderately engaged students who balanced content interaction with moderate session times and average completion rates [6]. Cluster 3 included students with low engagement, where interaction was limited, class time was shorter, and homework was done less often. These clusters offer a baseline

DOI: 10.54254/2753-8818/83/2025.19951

insight into how students engage, and this knowledge was used to create customized learning paths for each group.

3.3. Implementation Challenges

To predict academic performance and identify students potentially at risk, a decision tree model was developed based on the clustered engagement behaviors. Decision trees are highly interpretable models that use a tree-like structure to make predictions by splitting data based on feature values. In this study, the decision tree classified students into predicted performance categories, leveraging metrics from the engagement clusters. The model was trained on 70% of the dataset, with the remaining 30% used for cross-validation, allowing for accuracy assessment and model tuning. The model aimed to predict final academic outcomes (y) based on engagement metrics (x_i) and the resulting clusters, with the prediction function defined as:

$$y = f(x_1, x_2, x_3, \ldots, x_n) \tag{1}$$

where x_1 represents session duration, x_2 is assignment completion rate, x_3 denotes frequency of interaction with videos, and so forth. The tree splits on engagement metrics, with each branch representing a binary decision based on a threshold value [7]. Key nodes included splits on session duration and assignment completion rate, which emerged as the most significant predictors of academic success, highlighting their strong influence on learning outcomes. The final model achieved an accuracy of 82%, demonstrating its efficacy in categorizing students into performance brackets. This predictive framework offers online learning platforms a practical tool to flag students who might benefit from early interventions. By focusing on modifiable engagement behaviors, educators can design targeted support strategies, potentially improving academic outcomes and reducing dropout rates. This predictive approach thus represents an efficient, data-driven methodology for enhancing high school students' learning experiences on digital platforms [8].

4. Experimental Results

4.1. Cluster Characteristics

Each cluster exhibited distinct characteristics in terms of learning behavior and performance outcomes. Cluster 1, representing high-engagement students, showed an average session duration of 45 minutes and a quiz accuracy rate of 85%. This group consistently interacted with course materials and revisited resources prior to assessments, achieving high performance across subjects. Cluster 2, with moderate engagement, averaged 30 minutes per session and had a quiz accuracy of 70%, indicating a balanced yet less intensive study approach. Cluster 3, with low engagement, exhibited short sessions averaging 15 minutes and a quiz accuracy of 50%, highlighting gaps in content understanding and completion. A summary of these characteristics across the three clusters can be found in Table 1 [9].

In addition, the relationship between engagement levels and quiz accuracy is further illustrated in Figure 1, which visually represents the quiz accuracy rate across the clusters [10]. This chart underscores the positive correlation between engagement behaviors and learning outcomes, particularly as high-engagement students consistently outperform those with lower engagement levels.

DOI: 10.54254/2753-8818/83/2025.19951

Table 1: Experimental Results - Cluster Characteristics

Cluster	Average Session Duration (minutes)	Quiz Accuracy Rate (%)	Assignment Completion Rate (%)
Cluster 1 (High Engagement)	45	85	90
Cluster 2 (Moderate Engagement)	30	70	75
Cluster 3 (Low Engagement)	15	50	55

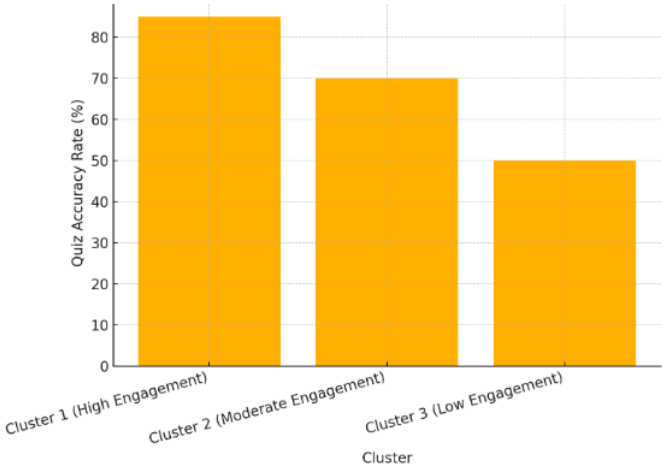

Figure 1: Quiz Accuracy Rate by Cluster

4.2. Personalized Pathway Construction

Personalized pathways were developed for each cluster to address their unique learning needs, optimizing engagement and learning outcomes. For Cluster 1, high-engagement students received advanced resources, including challenging exercises and extension projects. This additional content aimed to keep these students motivated and support deeper learning, aligning with their readiness for more complex material. Cluster 2, with moderate engagement, was provided with intermediate resources and self-assessment tools that reinforced core concepts without overwhelming them. Feedback on quizzes and assignments helped these students reflect on their progress, maintaining balanced engagement and building foundational skills. For Cluster 3, low-engagement students, the pathway emphasized foundational skills, increased scaffolding, and frequent reminders. Basic instructional materials were complemented by step-by-step guidance, while reminders prompted these students to re-engage and complete tasks. Early results show improved quiz scores in Cluster 3, indicating that this targeted support successfully closed engagement and understanding gaps for this group.

4.3. Predictive Model Performance

The decision tree model we used in this study accurately predicted high-performing students and low-performing students with accuracy values which clearly demonstrate the success of the engagement predictors used. Through the use of session time and content completion rates as decision nodes, the model correctly ranked students into predicted performance categories, reflecting the relevance of those measures for academic success. Time spent in the session, which is the biggest predictor, defines how long you are doing it, if at all, and this can affect your comprehension and retention. The students who took classes for more than a half an hour generally engaged with the material more frequently, and scores on quizzes and assignments correlated well [11]. Completion of material was another good

DOI: 10.54254/2753-8818/83/2025.19951

marker; the percentage of homework they completed revealed greater understanding, because the act of completing tasks reinforces learning by application. The interaction frequency, although a moderate predictor, added additional information on the level of engagement in a student but had less influence on classification accuracy of the model. Decision nodes in the model explain how all these variables interact to calculate a student's probability of performing well or poorly. For instance, high session length and high content completion rates consistently ranked students as high performers, while low session length and low content completion rates consistently ranked students as low performers. This model achieved an accuracy of 82% and was cross-validated with a second dataset to confirm its predictive power and usefulness.

5. Conclusion

The objective of this research is to use data mining and machine learning methods to track the behavior of high school students through online education platforms, so that we can learn trends, create personalized pathways, and forecast academic performance. We took engagement metrics including sessions, videos watched, and assignment completion rates from 1,000 high school students. We used a k-means clustering algorithm to break students into high, moderate and low engagement groups with different learning habits and academic performance patterns. Customised learning modules were then crafted for each cohort, tailoring content to maximise participation and facilitate learning. They also employed a decision tree model to predict performance which was 82% accurate with session length and assignment completion rates being among the best predictors. The results show that targeted interventions, rooted in engagement-based data, have the potential to positively impact learning experiences on digital platforms for high school students.

References

[1] *Baek, Clare, and Tenzin Doleck. "Educational data mining versus learning analytics: A review of publications from 2015 to 2019." Interactive Learning Environments 31.6 (2023): 3828-3850.*

[2] *Lampropoulos, Georgios. "Educational data mining and learning analytics in the 21st century." Encyclopedia of data science and machine learning. IGI Global, 2023. 1642-1651.*

[3] *Cerezo, Rebeca, et al. "Reviewing the differences between learning analytics and educational data mining: Towards educational data science." Computers in Human Behavior 154 (2024): 108155.*

[4] *Sghir, Nabila, Amina Adadi, and Mohammed Lahmer. "Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022)." Education and information technologies 28.7 (2023): 8299-8333.*

[5] *Ouyang, Fan, et al. "Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course." International Journal of Educational Technology in Higher Education 20.1 (2023): 4.*

[6] *Susnjak, Teo. "Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and ChatGPT." International Journal of Artificial Intelligence in Education 34.2 (2024): 452-482.*

[7] *Sailer, Michael, et al. "The End is the Beginning is the End: The closed-loop learning analytics framework." Computers in Human Behavior (2024): 108305.*

[8] *Saqr, Mohammed, and Sonsoles López-Pernas. "Learning analytics methods and tutorials: A practical guide using R." (2024): 736.*

[9] *Yan, Lixiang, Roberto Martinez-Maldonado, and Dragan Gasevic. "Generative artificial intelligence in learning analytics: Contextualising opportunities and challenges through the learning analytics cycle." Proceedings of the 14th Learning Analytics and Knowledge Conference. 2024.*

[10] *Vashishth, Tarun Kumar, et al. "AI-driven learning analytics for personalized feedback and assessment in higher education." Using traditional design methods to enhance AI-driven decision making. IGI Global, 2024. 206-230.*

[11] *Tzimas, Dimitrios E., and Stavros N. Demetriadis. "Impact of learning analytics guidance on student self-regulated learning skills, performance, and satisfaction: A mixed methods study." Education Sciences 14.1 (2024): 92.*