GAS Journal of Engineering and Technology (GASJET)

OPEN ACCESS

Volume 2, Issue 4, 2025 Journal Homepage: https://gaspublishers.com/gasjet-home/ ISSN: 3048-5800
Email: editor@gaspublishers.com

Data Mining in Education: Identifying At-Risk Students from LMS Logs

Brenden Donguines, France Einstein A. Baterna

State University of Northern Negros, Philippines

Received: 19.06.2025 | **Accepted:** 24.07.2025 | **Published:** 26.07.2025

***Corresponding Author:** Brenden Donguines

DOI: *10.5281/zenodo.16459908*

Abstract **Original Research Article**

The integration of Learning Management Systems (LMS) in educational institutions has revolutionized teaching and learning processes by enabling digital access, activity tracking, and performance monitoring. However, despite the growing availability of LMS data, many institutions struggle to identify students at risk of academic failure in a timely manner. This study explores the use of data mining techniques to analyze LMS log data and detect patterns indicative of at-risk student behaviour.

Using a dataset derived from a university LMS platform, the study focuses on key engagement metrics such as login frequency, time spent on the platform, assignment submission punctuality, quiz scores, and participation in discussion forums. Classification algorithms including Decision Trees and Naïve Bayes, as well as unsupervised clustering methods like k-Means, are employed to analyze the data and predict at-risk profiles.

The results reveal significant correlations between digital engagement and academic performance. Students with low login frequency and delayed submissions were more likely to fall into the at-risk category. The predictive models achieved an accuracy rate of over 85%, indicating the reliability of data mining techniques in educational analytics.

These findings underscore the potential of LMS log analysis as a proactive tool for educators and academic advisors. By identifying at-risk students early in the term, institutions can implement targeted interventions to improve learning outcomes and reduce dropout rates. This research contributes to the growing field of Educational Data Mining (EDM) and highlights the importance of data-driven decision-making in modern academic environments.

Keywords: At-Risk Students, Classification, Educational Data Mining, Learning Management System, LMS Logs, Student Performance.

Citation: Donguines, B., & Baterna, F. E. A. (2025). Data mining in education: Identifying at-risk students from LMS logs. *GAS Journal of Engineering and Technology (GASJET)*, *2*(4), [15-21].

INTRODUCTION

In recent years, the widespread adoption of Learning Management Systems (LMS) has transformed the educational landscape, offering flexible and accessible learning experiences. Platforms such as Moodle, Blackboard, and Google Classroom allow institutions to deliver content, monitor learner activity, and assess academic performance through a centralized digital environment. Alongside this shift, LMS platforms generate large volumes of data that capture every interaction between students and the system ranging from logins and assignment submissions to quiz attempts and forum participation.

According to Al-Rawahna (2020), data mining is essential across various sectors because accessing and analyzing large volumes of data requires both time and a high level of accuracy. In the field of higher education, the potential impact of data mining on student learning processes and outcomes is becoming increasingly evident. The data generated by LMS platforms presents a valuable opportunity to apply data mining techniques to uncover hidden patterns and trends in student behaviour. Data mining, a core component of knowledge discovery in databases (KDD), involves extracting meaningful information from large datasets to support decision-making (e.g., Han, Pei, & Kamber, 2011). In educational settings, this process referred to as Educational Data Mining (EDM) has gained traction as a means of enhancing teaching and learning through evidence-based insights. In the era of social information, the application of data mining technology to the evaluation of teaching quality in higher vocational colleges has gradually become an important way for higher vocational colleges to improve the quality of teaching and talent training

(e.g., Chen, Liu, Zheng, 2022)

Identifying at-risk students early in the academic cycle is critical, as it allows educators to implement timely interventions that improve student retention and success (e.g., Romero & Ventura, 2020). Traditional indicators of academic risk, such as low grades or chronic absenteeism, often appear too late to mitigate learning challenges effectively. In contrast, analyzing LMS logs can offer real-time behavioural insights, including inactivity, low participation, delayed submissions, or poor assessment performance all of which can serve as early warning signs.

By applying data mining algorithms such as classification, clustering, and regression researchers can build predictive models that flag students at risk based on their LMS usage patterns (e.g., Baker & Inventado, 2014). These models support institutional efforts to enhance student outcomes and contribute to the growing field of learning analytics, which focuses on leveraging educational data to inform instruction and promote academic achievement.

OBJECTIVES OF THE STUDY

Our research aims to identify at-risk students by analyzing Learning Management System (LMS) log data using data mining techniques. It exactly focuses on Key Behavioural Indicators (KBIs) like login rate, their time spent on the platform, their punctuality on assignment submission, quiz scores, and forum participation, which are considered potential predictors of academic performance. It will measure the significance of these variables and evaluate the correctness, accuracy, recall, and F1-score of predictive models developed using classification algorithms like Decision Tree and Naïve Bayes, and of course the clustering methods like k-Means. These methods will be applied to historical LMS data to classify patterns and students according to academic risk. The research is possible within the semester, with a clear timeline for data collection, model development, analysis and validation. The results are intended to provide related insights that support early intervention strategies for teachers and academic institutions. By the end of the semester, the study aims to deliver a working predictive model, practical recommendations for academic support, and a scholarly contribution to the field of Educational Data Mining.

MATERIALS AND METHODS

A. Data Source and Collection

This research made use of anonymized log records sourced from a university's Learning Management System (LMS), with a particular emphasis on undergraduate students enrolled in general education and core subjects during the 2023–2024 academic year. The dataset captured essential indicators of student engagement, including:

- Frequency of logins (daily or weekly).
- Duration of time spent on the platform.
- Timestamps of assignment submissions.
- Scores from quizzes and exams.
- Level of participation in online discussion forums.

Strict data privacy guidelines were followed throughout the process. Ethical approval was secured from the institution's research ethics board, and all personally identifiable information (PII) was either removed or anonymized to safeguard student identities.

B. Data Pre-processing

To ensure the data was suitable for analysis, a series of pre-processing steps were performed, including:

- Addressing missing data by applying mean substitution for numerical variables and mode substitution for categorical ones.
- Normalizing time-related data (e.g., total hours spent) to prevent scale-related distortions.
- Converting categorical variables like forum participation into numerical values.
- Grouping activity logs by student ID to create complete behavioral profiles for each participant.

After processing, the dataset included more than 1,000 student entries, each expressed as a structured set of LMS engagement attributes.

C. Feature Selection

To identify which variables most significantly influenced academic performance, correlation analysis and exploratory data analysis (EDA) were conducted. Any features that demonstrated weak relationships with final grades or exhibited high interdependence with other variables were excluded to enhance model reliability and prevent over fitting.

The final set of selected features included:

- Frequency of logins.
- Average session length.
- Timeliness in submitting assignments.
- Mean quiz scores.
- Number of discussion forum posts.

D. Model Development

Two main types of data mining techniques were applied:

1. **Classification Techniques:** Decision Tree (using the C4.5 algorithm): Chosen for its interpretability and effectiveness in handling categorical data.

Naïve Bayes: Selected for its simplicity and suitability for probabilistic prediction based on feature independence.

2. **Clustering Techniques:** K-Means Clustering: Applied to group students based on similar behavioural patterns, aiding in the visualization of at-risk versus high-performing profiles.

E. Model Evaluation

To validate the predictive models, the dataset was divided into training (70%) and testing (30%) sets. Model performance was evaluated using the following metrics:

- **Accuracy:** Overall percentage of correct

classifications.

- **Precision:** Proportion of true positive predictions among all positive predictions.
- Recall (Sensitivity): Proportion of actual at-risk students correctly identified.
- **F1 Score:** Harmonic mean of precision and recall, balancing false positives and false negatives. Cross-validation (10-fold) was also employed to ensure model generalizability and robustness.

F. Tools and Software

The following tools were used in the study:

- **Python** (with libraries such as pandas, scikit-learn, and matplotlib) for data analysis and machine learning. Python is considered the most suited for these purposes, especially in engineering context dependent or temporal features compared to Excel and Sheets (e.g., Salihoun, 2020).
- **WEKA** for additional model testing and visualization. WEKA is free and open source software including a set of algorithms related to machine learning. It offers tools for data mining tasks such as regression, classification, association, rules mining, clustering, and visualization (e.g., Singhal, Jena, 2013).
- **Microsoft Excel** is for initial data inspection. The visual inspection of data and features created in Excel it is easier than in Jupyter (e.g., Salihoun, 2020).
- **Jupyter Notebook** is the main development environment. It is a web-based interactive computational environment having a useful feature that allows creating and sharing document including data cleaning and transformation, numerical simulation, data visualization, statistical modelling, machine learning, etc. (e.g., Kluyver et al., 2016).

SAMPLING TECHNIQUE

To ensure equitable representation of students across varying academic performance levels, the study utilized a stratified random sampling approach. The initial dataset included LMS log records from more than 2,500 students spanning different departments and academic year levels.

To prevent imbalances that might bias the predictive models, students were categorized into three performance groups based on their final grades:

- **High-performing:** Grades of 90 and above.
- **Average:** Grades ranging from 75 to 89.
- **At-risk:** Grades below 75.

A proportionate number of students from each group were randomly chosen, resulting in a total of 1,000 students. This method ensured each performance category was fairly represented during both model training and evaluation stages.

Data Collection Procedure

The data collection process adhered strictly to institutional ethics guidelines and was designed to protect student privacy. Approval was obtained from the university's IT department and academic affairs office to access anonymized LMS logs for academic research purposes. The research team collaborated with system administrators to retrieve raw log data from the university's LMS, covering both the first and second semesters of the 2023–2024 academic years.

The extraction process used secure, read-only access and was limited to activity logs directly related to student academic engagement. Inclusion criteria were set to filter only active undergraduate students enrolled in general education and core courses that had complete grade records and LMS usage data for the entire academic period.

To ensure privacy, all personal identifiers such as names and student ID numbers were replaced with randomly generated codes, complying with RA 10173, the Data Privacy Act of 2012. The resulting dataset included time stamped logs of key LMS activities like logins, quiz participation, and assignment submissions, organized in structured formats for pre-processing and analysis.

All data-handling procedures were thoroughly documented and reviewed by the university's research ethics board to maintain confidentiality and uphold academic research standards.

DATA ANALYSIS

This stage focused on interpreting patterns, correlations, and classification outcomes derived from LMS log data using selected data mining methods. The objective was to pinpoint behavioral indicators that consistently signal at-risk students based on their LMS activity patterns. Behavioral intention is an important predictor of student behavior that varies between different behavioral, control, and normative beliefs on the desired behavior (e.g., Križanić, 2020). The analysis step identifies the existing interesting patterns, which can be displayed for a better visualization (Han et al., 2011)

A. Descriptive Statistics

Preliminary statistical analysis across the 1,000 sampled students revealed the following averages:

- **Login Frequency:** 3.4 logins per week.
- **Session Duration:** Average of 27 minutes per session.
- **Assignment Punctuality:** 82% of assignments submitted before deadlines.
- **Quiz Performance:** Average score of 76%.
- **Discussion Participation:** 1.8 forum posts per week.

Students identified as at-risk (final grades below 75) consistently showed lower engagement across these indicators.

B. Correlation Analysis

Pearson correlation was used to examine how specific LMS activities relate to students' final grades. Key findings include:

- **Login Frequency:** Strong positive correlation (r = 0.64)
- **Assignment Timeliness:** Moderate positive correlation (r = 0.57)

- **Quiz Scores:** Strongest correlation observed (r = 0.71)
- **Forum Engagement:** Weak to moderate positive correlation (r = 0.41)
- **Session Duration:** Moderate positive correlation (r = 0.53)

These results informed the selection of relevant features for building accurate prediction models.

C. Classification Model Outcomes

1. Decision Tree (C4.5 Algorithm)

- **Accuracy:** 87.2%
- **Precision:** 84.5%
- **Recall:** 85.3%
- **F1 Score:** 84.9%

The decision tree model offered clear, interpretable rules. For instance, students who logged in less than twice per week and submitted over 40% of assignments late were frequently classified as at-risk.

2. Naïve Bayes Model

- **Accuracy:** 81.6%
- **Precision:** 78.9%
- **Recall:** 80.2%
- **F1 Score:** 79.5%

Although slightly less accurate, the Naïve Bayes classifier provided a lightweight and efficient alternative, making it well-suited for integration into real-time risk detection systems.

D. Clustering Results (k-Means)

Using k = 3 to reflect the predefined academic performance groups (high-performing, average, and at-risk), k-Means clustering effectively categorized students into distinct engagement profiles:

- **Cluster 1**: High-performing students—frequent logins, timely submissions, and high quiz results
- **Cluster 2**: Average students—moderate activity and performance
- **Cluster 3:** At-risk students—low engagement, late submissions, and low quiz scores

This clustering visually confirmed the segmentation and also revealed transitional students—those whose behaviors placed them between average and at-risk classifications.

E. Cross-Validation

Ten-fold cross-validation confirmed the robustness of both classification models. The standard deviation of accuracy across folds remained under 3% for both models, indicating consistent predictive performance across different data splits.

RESULTS AND DISCUSSION

The implementation of data mining models in this research has offered valuable insights into student behavior within a Learning Management System (LMS) and its connection to academic performance. Through classification and clustering methods, the study effectively identified students at academic risk by analyzing their LMS engagement patterns. The results demonstrate both the predictive strength and practical use of these methods in educational environments.

A. Effectiveness of Predictive Models

Among the models tested, the Decision Tree emerged as the top performer with an accuracy of 87.2%, surpassing the Naïve Bayes model, which recorded 81.6% accuracy. The Decision Tree's recall rate of 85.3% indicates its strong ability to detect at-risk students essential for timely interventions. These findings are consistent with earlier studies (e.g., Romero & Ventura, 2020), which favor decision trees due to their clarity and ease of interpretation.

While Naïve Bayes was slightly less accurate, it offered faster processing and simplicity, making it suitable for real-time monitoring systems or environments with limited computing capacity. Choosing the right model depends on balancing accuracy, interpretability, and computational efficiency for the institution's specific needs.

B. Key Behavioral Risk Indicators

Analysis of student interaction data in the LMS highlighted recurring behavior among those with low academic performance:

- Infrequent logins (typically fewer than twice weekly).
- High percentage of late assignment submissions (over 40%).
- Below-average quiz/exam scores (often under 76%).
- Minimal participation in forums (usually less than one post weekly).

These indicators were statistically validated and support existing literature (e.g., Baker & Inventado, 2014), confirming that basic digital engagement metrics can reliably serve as early warning signs for academic risk.

C. Insights from Cluster Analysis

The **k-Means** clustering (with k = 3) visually demonstrated clear distinctions among student engagement levels. Students in Cluster 3, classified as at-risk, consistently displayed low levels across all LMS activity measures. This clustering reinforced the classification outcomes and showcased how unsupervised learning can uncover patterns that may not emerge through supervised models alone.

Additionally, the identification of borderline or transitional students those hovering between average and at-risk suggests the potential for dynamic risk tracking systems. These systems could adapt to behavioral changes over time, offering timely updates to risk assessments and enabling more targeted interventions.

D. Educational Implications

This study highlights the valuable role of **predictive analytics** in enhancing student support through integration into **Learning Management Systems (LMS)**. By leveraging real-

time and historical data, schools can shift from reactive responses to proactive intervention particularly for students at risk of academic failure or disengagement.

1. Value of Predictive Analytics in Education

By analyzing attendance, participation, assignment submissions, and grades, predictive tools can:

- **Trigger automated alerts** when students show signs of declining performance (e.g., missed deadlines, low engagement).
- **Support data-driven decisions**, allowing educators to intervene effectively based on reliable insights.
- **Enable scalable monitoring**, helping staff oversee large cohorts without manual tracking.

2. Personalized Student Support

Early risk detection allows for tailored interventions, such as:

- **Academic guidance** in study skills and time management.
- **Customized learning plans** or tutoring for academic recovery.
- **Mental health referrals**, if behavioral data suggests emotional distress.

This approach enhances engagement by showing students that help is available and specific to their needs.

3. Early Access to Support Services

Predictive alerts can prompt timely referrals to:

- Remedial workshops
- Peer tutoring
- Writing/math labs
- Additional instructional sessions

This proactive strategy increases the likelihood of student improvement before failure occurs.

4. Smarter Resource Allocation

Analytics also help institutions optimize resources by:

- Assigning **mentors** to students who need it most
- **Investing strategically** in programs that boost retention
- Identifying high-risk **departments or courses** for targeted support

5. A Four-Step Framework for Implementation

To effectively apply predictive analytics, a structured approach is recommended:

a. **Monitor:** Track LMS logins, grades, submissions, and participation.
b. **Detect:** Use models to identify at-risk students based on data trends.
c. **Intervene:** Provide personalized outreach and academic support.
d. **Evaluate:** Review outcomes and refine strategies using feedback and new data.

6. Key Benefits

When implemented effectively, predictive analytics lead to:

- **Higher retention** through early intervention
- **Better academic performance** via personalized support
- **Increased student satisfaction** and trust
- **A stronger data-informed culture** among educators and administrators

E. Study Limitations and Future Directions

While the predictive models used in this study demonstrated promising outcomes, several limitations must be acknowledged. These constraints affect the generalizability and completeness of the findings and offer valuable insights for future research and application.

Single-Institution Data: The analysis was based on data from only one university, which limits how widely the findings can be applied. Different institutions may yield different outcomes due to variations in student demographics, teaching methods, and academic policies. Future studies should involve multiple universities for broader applicability.

Limited Course Coverage: only general education and core undergraduate courses were included, excluding major-specific or advanced courses. Future research should explore a wider range of disciplines and academic levels.

Lack of Non-Academic Variables: The study focused solely on LMS data and did not account for factors like student motivation, socio-economic status, or personal challenges, which can significantly impact performance. Future work should integrate both academic and non-academic data for a more complete view.

CONCLUSION & RECOMMENDATION

This research highlighted the value of using data mining techniques to effectively identify students at academic risk through their interactions within a Learning Management System (LMS). By examining behavioral indicators such as frequency of logins, time spent per session, punctuality in submitting assignments, quiz performance, and participation in forums, the study confirmed a strong link between online engagement and academic outcomes. These techniques can benefit various fields through different objectives, such as extracting patterns, predicting behavior, or describing trends (e.g., Alyahyan, Düştegör, 2020). Mining big data in education challenges not only how we prepare education researchers, but also what kinds of research practices we engage in (e.g., Pardos et al., 2020).

The use of classification algorithms Decision Tree and Naïve Bayes alongside k-Means clustering resulted in predictive models with notable accuracy. The Decision Tree model performed particularly well, achieving an 87.2% accuracy rate. The clustering analysis further emphasized the clear behavioral distinctions between high-achieving, average, and at-risk students, offering deeper understanding into student

engagement patterns.

RECOMMENDATIONS

Based on the results of the study, the following recommendations are proposed:

Implement Early Warning Systems in LMS Platforms

Educational institutions should integrate predictive analytics tools within their LMS to monitor student activity in real-time. These systems can automatically flag at-risk students based on defined behavioural thresholds and notify instructors or advisers.

Train Faculty and Advisers in Data-Driven Intervention

Teachers and academic advisers should be trained on how to interpret LMS analytics reports and use them to provide timely support. Workshops and guides on educational data mining can help bridge the gap between technical insights and pedagogical action.

Enhance Student Engagement through Targeted Strategies

Institutions should develop targeted engagement strategies such as reminders for low-login students, encouragement for forum participation, or flexible submission options to re-engage those showing early signs of academic risk.

Periodically Evaluate and Refine Predictive Models

Predictive models must be regularly updated and validated with new data to maintain accuracy. Including more variables, such as attendance, psychological factors, or socio-economic indicators, may improve prediction capabilities.

Promote Ethical Use of Student Data

Institutions must continue to uphold data privacy standards and ensure that predictive analytics are used ethically. Transparency with students regarding how their data is used, along with consent-based mechanisms, is essential for trust and compliance.

Expand Research Scope for Broader Insights

Future studies should include a wider range of academic programs, institutions, and longitudinal data to improve the generalizability of the findings. Exploring the use of deep learning or hybrid models may also provide deeper insights into complex learning behaviours.

REFERENCES

Al-Rawahnaa, A. (2020). Data mining for Education Sector, a proposed concept. Journal of Applied Data Sciences, 1(1), 1–10. https://doi.org/10.47738/jads.v1i1.6

Baker, R. S. J. d., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), Learning analytics: From research to practice (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4

Han, J., Kamber, M., & Pei, J. (2011). Data Mining : Concepts and Techniques 3rd edition Ed. 3. In Elsevier eBooks. https://www.scholartext.com/book/88809627?_locale=fr

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355. https://doi.org/10.1002/widm.1355

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining Big data in Education: Affordances and challenges. Review of Research in Education, 44(1), 130–160. https://doi.org/10.3102/0091732x20903304

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. International Journal of Educational Technology in Higher Education, 17(1). https://doi.org/10.1186/s41239-020-0177-7

Salihoun, M. (2020). State of art of data mining and learning analytics tools in higher education. International Journal of Emerging Technologies in Learning (iJET), 15(21), 58. https://doi.org/10.3991/ijet.v15i21.16435

Singhal, S., & Jena, M. (2013). A study on WEKA tool for data preprocessing, classification and clustering. International Journal of Innovative technology and exploring engineering (IJItee), 2(6), 250-253.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., . & Ivanov, P. (2016, May). Jupyter Notebooks-a publishing format for reproducible computational workflows. In ELPUB (pp. 87-90). [15']

Križanić, S. (2020). Educational data mining using cluster analysis and decision tree technique: A case study. International Journal of Engineering Business Management, 12, 184797902090867.
https://doi.org/10.1177/1847979020908675

Chen, B., Liu, Y., & Zheng, J. (2022). Using data mining approach for student satisfaction with teaching quality in high vocation education. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.746558

Arnold, K. E., & Pistilli, M. D. (2012). *Course signals at Purdue: Using learning analytics to increase student success.* In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12), pp. 267–270. https://doi.org/10.1145/2330601.2330666

Slade, S., & Prinsloo, P. (2013). *Learning analytics: Ethical*

issues and dilemmas. American Behavioral Scientist, 57(10), 1510–1529. https://doi.org/10.1177/0002764213479366

Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013). *Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment.* https://oro.open.ac.uk/36574/