

A Systematic Review of Random Forest-Based Decision-Support Systems in Educational Management

Teodora – Mihaela SPIRIDON – MOCIOACĂ

University of Science and Technology POLITEHNICA Bucharest, Bucharest, Romania

ORCID ID: <https://orcid.org/0009-0002-6431-1556>

e-mail address: teodora.spiridon@upb.ro

Roxana-Adriana PUIU*

University of Science and Technology POLITEHNICA Bucharest, Bucharest, Romania

ORCID ID: <https://orcid.org/0009-0000-6072-3700>

e-mail address: roxana.puiu@upb.ro

Romina-Cătălina PASCARU

University of Science and Technology POLITEHNICA Bucharest, Bucharest, Romania

ORCID ID: <https://orcid.org/0009-0003-8187-5174>

e-mail address: romina.pascaru@upb.ro

Laura-Florentina BOANȚĂ

University of Science and Technology POLITEHNICA Bucharest, Bucharest, Romania

ORCID ID: <https://orcid.org/0000-0002-6627-3007>

e-mail address: laura.boanta@upb.ro

*Corresponding author, e-mail address: roxana.puiu@upb.ro

Abstract. An increasing number of educational institutions are using predictive analytics to help with strategic decision-making and programs that help students succeed. This paper looks into the most recent improvements in using supervised learning models, specifically Random Forest classifiers, to keep an eye on behavioral trends in digital learning settings. We focus on studies that use structured data exported from Learning Management Systems (LMS), with a concentration on CSV-based processing procedures because they are easy to use and understand. The research brings together several uses of predictive logic in schools, looking at how well the algorithms work, how easy they are to use, and how they were designed. This review provides a theoretical basis for building smart, user-friendly educational tools that can adapt to students' activity profiles by establishing consistent modeling patterns and ways to make them understandable.

Keywords: Behavioral trend modeling, Data-informed academic support, Supervised learning in education, CSV-format workflow design, Algorithmic interpretability

How to cite: SPIRIDON – MOCIOACĂ, T. – M., PUIU, R., PASCARU, R.-C. ., & BOANȚĂ, L.-F. . A Systematic Review of Random Forest-Based Decision-Support Systems in Educational Management. *International Conference of Management and Industrial Engineering*, 12. <https://doi.org/10.56177/12icmie2025.148>

Introduction

An increasing number of educational institutions are using data-driven methods to make decisions and improve student results. Learning analytics is at the center of this transition. It is the process of systematically gathering, evaluating, and reporting data about learners to improve learning environments and results [1]. When students use digital tools like Learning Management Systems (LMS), they leave behind detailed logs of their activity and performance, which can be used to get educational insight [2].

Learning Management Systems (LMS) are important for delivering courses and keeping track of students. They collect a lot of different types of interaction data, like logins, content access, forum posts, and assignment activity. These indications from LMS are very helpful for making prediction frameworks that can find students who are not engaged or are at danger early on [2],[3].

Supervised machine learning, especially Random Forest classifiers, have been used more and more on these datasets because they are robust, can handle high-dimensional data, and have built-in feature importance measures [4]. In educational settings, Random Forest models have shown to be quite good at making predictions, such as predicting student involvement and academic performance, as well as modeling student satisfaction [5]. More newer uses, such as interpretable ML employing LMS data, give us even more insight into how students behave [6].

The main objective of this study is to systematically synthesize research on Random Forest-based predictive models for educational decision support, focusing on methodological consistency and interpretability.

Literature Review

Recent research has demonstrated a growing interest in predictive analytics for educational management. Building upon prior reviews by Romero and Ventura [1], this section outlines key developments in applying supervised learning—particularly Random Forest models—to Learning Management System (LMS) data. The focus is on methodological robustness, interpretability, and the balance between theoretical rigor and practical implementation.

The necessity for scalable analytics solutions in institutions, along with the availability of data (via CSV or exportable LMS logs) and strong classification models, shows how important it is to have simplified, theory-driven tools. Studies have suggested frameworks that make it possible to find kids who are at risk using controlled workflows, frequently with an accuracy of over 90% [6],[7].

Despite this progress, there remains a need for accessible, modular analytics tools that non-technical educators and administrators can adopt without extensive infrastructure. Understanding methodological design patterns, interpretability strategies, and usability trade-offs is essential to bridge the gap between theoretical insight and practical tool deployment [9].

This literature review consolidates contemporary empirical and theoretical contributions in the use of supervised learning—especially Random Forest models—over structured LMS interaction datasets. By synthesizing approaches that prioritize usability, interpretability, and empirical performance, this review furnishes a rigorous foundation for the development of user-centered, lightweight analytic tools in educational settings.

Prior works differ in their scope and methodological rigor: while some emphasize algorithmic performance [5], others highlight interpretability and educator usability [10], [29]. This contrast reveals a growing trend toward balancing technical sophistication with human-centered design in educational analytics.

In 2021, Saleem et al. [6] presents a decision-support system that uses ML classifiers (Random Forest, AdaBoost, Bagging) on e-learning participation, assignment completion, and interaction logs (Kalboard 360 LMS). The system achieves prediction accuracies over 90% and provides teachers with intelligent recommendations for interventions. The workflow aligns with real-world scenarios where data is collected from CSV files and used for automated decision-making.

Based on these findings, the next section details the methodological framework adopted in this systematic review.

Methodology

This review follows a structured process for identifying, selecting, and analyzing scholarly works that investigate the use of supervised machine learning—particularly Random Forest models—in processing LMS-derived interaction data. The goal was to gather a representative set of studies covering both methodological and applied aspects of predictive analytics in education.

Search strategy

The literature search was conducted across multiple scholarly databases, including Google Scholar, IEEE Xplore, Scopus, SpringerLink, and MDPI Journals. Search strings combined domain-specific terms, model types, and data formats. Representative queries included:

- “Random Forest” AND “Learning Management System” AND “student engagement”
- “supervised learning” AND “LMS data” AND “predictive analytics”
- “educational decision support” AND “CSV-based data”
- “machine learning” AND “academic performance prediction” AND “log data”

Searches were restricted to publications in English, from 2015 to 2025, to ensure coverage of recent methodological advances and real-world implementations.

Inclusion and exclusion criteria

Studies were included if they met all the following criteria:

- Applied supervised machine learning with explicit use of Random Forest or compared it with other classifiers.
- Utilized real interaction data from LMS platforms (e.g., Moodle, Google Classroom, Canvas), preferably in CSV or equivalent tabular format.
- Reported empirical results with evaluation metrics such as accuracy, F1-score, precision, or recall.
- Were peer-reviewed or published in indexed conference proceedings.

Studies were excluded if they:

- Focused solely on unsupervised learning or clustering without engagement prediction.
- Relied exclusively on synthetic datasets unrelated to LMS behavior.
- Were conceptual papers without implemented models or performance evaluation.

Selection process

The initial search returned 228 records. After removing duplicates and screening titles and abstracts, 61 papers were kept for a full-text review. Using the inclusion and exclusion criteria brought this group down to 32 studies for synthesis. These studies reflect both breadth (covering multiple LMS platforms and educational levels) and depth (detailing preprocessing workflows, model tuning, and interpretation techniques).

Data extraction and analysis

For each study, the following elements were extracted systematically:

- Educational context (platform used, learner demographics).
- Dataset properties (source, format, size, features).
- Preprocessing techniques (encoding, balancing, feature selection).
- Modeling details (Random Forest parameters, benchmark models).
- Performance metrics (accuracy, precision, recall, F1-score).
- Interpretability features (feature importance, explainability methods).
- Usability aspects (deployment context, user interface, accessibility).

Findings were organized in a narrative synthesis complemented by comparative tables, enabling direct observation of trends, methodological patterns, and practical design considerations across the reviewed literature.

To ensure analytical consistency, the studies were cross-validated against inclusion criteria emphasizing both methodological soundness and educational applicability. Random Forest was selected as the focal algorithm due to its ensemble nature, interpretability through feature importance, and proven generalization capacity in prior meta-analyses [4], [21].

Results and further discussions

The systematic review identified 32 relevant studies, spanning multiple educational contexts, LMS platforms, and predictive modeling strategies. Analysis of these studies revealed recurrent methodological patterns and several noteworthy trends.

Prevalence of Random Forest in educational predictive analytics

Random Forest was the most frequently used classifier across the reviewed literature, cited in more than 80% of the selected works. Studies reported predictive accuracies generally ranging from 85% to 96%, consistently outperforming simpler models such as Decision Trees or Naïve Bayes ([4], [5], [6], [9], [10], [11]).

For example, Jawad et al. [5] achieved 94.8% accuracy when predicting course completion likelihood from LMS logs. Other works, such as Luo et al. [12], combined RF with spatiotemporal features to enhance early risk detection, confirming findings from Biau & Scornet [4] regarding the algorithm's robustness.

Use of CSV-based and exportable LMS data

A significant proportion of the studies relied on CSV-formatted exports or other tabular datasets derived from Moodle, Google Classroom, Canvas, or institutional LMS platforms [1], [5], [6], [7], [9], [10], [11], [13], [14],

[15], [16], [17], [18], [19], [20], [21], [22]. This preference reflects the accessibility and interoperability of CSV data, especially in institutions with limited IT infrastructure.

Santos [13] and Chen et al. [14] highlighted that such formats simplify preprocessing workflows (encoding, normalization, balancing), while Waheed et al. [20] demonstrated that they enable rapid deployment of prediction models without the need for complex database integration.

Feature engineering and importance

Feature selection emerged as a critical factor in improving both performance and interpretability. The most frequently used predictors across the reviewed studies included engagement metrics such as login frequency, resource views, and forum activity [1], [14], [17], [23], [24]. Performance indicators like assignment completion rates and quiz or exam scores [5], [17], [21], [24] were also commonly employed to measure academic outcomes. In addition, several studies incorporated behavioral or qualitative signals—such as teacher feedback, sentiment analysis, and peer interaction quality—to capture affective and social aspects of learning [15], [25]. Random Forest’s built-in feature importance ranking was often leveraged to identify the most influential variables, as observed in Jiao et al. [10] and Nachouki et al. [11], supporting the theoretical insights of Biau & Scornet [4].

Addressing class imbalance

Several studies identified the problem of class imbalance, where the number of engaged learners significantly exceeded that of disengaged ones [24]. To mitigate this challenge, different resampling and balancing strategies were proposed to improve the representativeness of training data. Building on these approaches, Brahim et al. [24] demonstrated that the use of resampling techniques can enhance the detection of at-risk or underperforming learners, thereby improving model sensitivity and overall educational relevance.

Integration into Decision-Support Systems

While many studies focused on model performance, fewer explored integration into operational decision-support tools ([9], [13], [15], [16], [25]).

Johnston et al. [7] and Angeioplastis et al. [12] demonstrated interpretable dashboards for risk monitoring in Moodle. Joarder et al. [29] stressed that a “flexible user interface” increases educator adoption, while Alnasyan et al. [25] highlighted that even deep learning-based pipelines benefit from visual explainability features.

Comparative approaches and hybrid models

Although RF was dominant, some studies compared it with or embedded it into ensemble or hybrid approaches ([9], [10], [11], [21]).

For example, Alyahyan & Düşteğör [21] presented systematic reviews where RF appeared consistently among the top-performing algorithms.

Best practices and research opportunities

The synthesis of reviewed literature highlights several best practices that can guide future research and implementation. Scholars emphasize the early incorporation of interpretability tools such as SHAP and LIME to ensure transparency in predictive models [4], [10], [15]. Continuous or iterative model updates are also recommended to adapt to changing learner behaviors and evolving educational environments [6], [13]. Moreover, combining quantitative LMS metrics with qualitative feedback—such as sentiment analysis or teacher input—has been shown to enrich prediction accuracy and contextual insight [15], [25]. Another recurring recommendation is the use of lightweight, CSV-based workflows to enable scalability and ease of deployment in low-resource institutional contexts [1], [7], [20].

Despite notable progress, research opportunities remain in developing cross-institutional datasets [21], improving real-time processing capabilities [13], [25], and integrating predictive models into closed-loop intervention systems that provide actionable insights to educators [9].

Conclusion

This systematic review examined a corpus of 32 peer-reviewed studies published between 2015 and 2025 that applied supervised machine learning—particularly Random Forest classifiers—to interaction data derived from Learning Management Systems. The analysis revealed that Random Forest consistently delivers strong predictive performance, often surpassing alternative models in accuracy, robustness, and feature interpretability.

A substantial proportion of the reviewed works demonstrated the practical value of using CSV-formatted or other tabular LMS exports, which offer accessibility, interoperability, and low-infrastructure deployment potential. The studies also highlighted the importance of feature engineering, class imbalance handling, and the integration of interpretability methods to ensure actionable insights for educators and administrators.

While several implementations successfully bridged predictive modeling and decision-support interfaces, broader adoption remains limited by technical deployment challenges and the need for adaptable, cross-institutional datasets. Promising directions for future research include incorporating richer qualitative feedback, expanding interoperability across diverse LMS platforms, and embedding predictive analytics into closed-loop intervention systems that directly support learner success.

By consolidating methodological patterns, performance benchmarks, and usability considerations, this review provides a theoretical and empirical foundation for the design of intelligent, user-oriented educational technologies. The synthesis aims to guide both researchers and practitioners in developing scalable, transparent, and context-sensitive decision-support solutions that can enhance data-driven educational management in the coming years.

References

- [1] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10 (3). DOI: <https://doi.org/10.1002/widm.1355>
- [2] Ahmadi, G., Mohammadi, A., Asadzandi, S., Shah, M., & Mojtahedzadeh, R. (2023). What are the indicators of student engagement in learning management systems? a systematized review of the literature. *International Review of Research in Open and Distributed Learning*, 24(1), 117-136. DOI: <https://doi.org/10.19173/irrodl.v24i1.6453>
- [3] Hamadi, H., Tafili, A., Kates, F. R., Larson, S. A., Ellison, C., & Song, J. (2023). Exploring an innovative approach to enhance discussion board engagement. *TechTrends*, 67(4), 741-751. DOI: <https://doi.org/10.1007/s11528-023-00850-0>
- [4] Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, 25(2), 197-227. DOI: <https://doi.org/10.1007/s11749-016-0481-7>
- [5] Jawad, K., Shah, M. A., & Tahir, M. (2022). Students' academic performance and engagement prediction in a virtual learning environment using random forest with data balancing. *Sustainability*, 14(22), 14795. DOI: <https://doi.org/10.3390/su142214795>
- [6] Guo, Y., Gunay, C., Tangirala, S., Kerven, D., Jin, W., Savage, J. C., & Lee, S. (2022). Identifying critical lms features for predicting at-risk students. *arXiv preprint arXiv:2204.13700*. DOI: <https://doi.org/10.48550/arXiv.2204.13700>
- [7] Pan, Z., Biegley, L., Taylor, A., & Zheng, H. (2024). A systematic review of learning analytics: Incorporated instructional interventions on learning management systems. *Journal of Learning Analytics*, 11(2), 52-72. DOI: <https://doi.org/10.18608/jla.2023.8093>
- [8] Joarder, S., Chatti, M. A., Mirhashemi, S., & Ain, Q. U. (2023). Towards a Flexible User Interface for'Quick and Dirty'Learning Analytics Indicator Design. *arXiv preprint arXiv:2304.01711*. DOI: <https://doi.org/10.48550/arXiv.2304.01711>
- [9] Saleem, F., Ullah, Z., Fakieh, B., & Kateb, F. (2021). Intelligent decision support system for predicting student's e-learning performance using ensemble machine learning. *Mathematics*, 9(17), 2078. DOI: <https://doi.org/10.3390/math9172078>
- [10] Jiao, P., Ouyang, F., Zhang, Q., & Alavi, A. H. (2022). Artificial intelligence-enabled prediction model of student academic performance in online engineering education. *Artificial Intelligence Review*, 55(8), 6321-6344. DOI: <https://doi.org/10.1007/s10462-022-10155-y>
- [11] Nachouki, M., Mohamed, E. A., Mehdi, R., & Abou Naaj, M. (2023). Student course grade prediction using the random forest algorithm: Analysis of predictors' importance. *Trends in Neuroscience and Education*, 33, 100214. DOI: <https://doi.org/10.1016/j.tine.2023.100214>
- [12] Luo, Z., Mai, J., Feng, C., Kong, D., Liu, J., Ding, Y., ... & Zhu, Z. (2024). A method for prediction and analysis of student performance that combines multi-dimensional features of time and space. *Mathematics*, 12(22), 3597. DOI: <https://doi.org/10.3390/math12223597>
- [13] Santos, R. M., & Henriques, R. (2023). Accurate, timely, and portable: Course-agnostic early prediction of student performance from LMS logs. *Computers and Education: Artificial Intelligence*, 5, 100175. DOI: <https://doi.org/10.1016/j.caeai.2023.100175>

- [14] Chen, F., & Cui, Y. (2020). Utilizing Student Time Series Behaviour in Learning Management Systems for Early Prediction of Course Performance. *Journal of Learning Analytics*, 7(2), 1-17.
- [15] Azizah, Z., Ohyama, T., Zhao, X., Ohkawa, Y., & Mitsuishi, T. (2024). Predicting at-risk students in the early stage of a blended learning course via machine learning using limited data. *Computers and Education: Artificial Intelligence*, 7, 100261. DOI: <https://doi.org/10.1016/j.caeai.2024.100261>
- [16] Angeioplastis, A., Aliprantis, J., Konstantakis, M., & Tsimpiris, A. (2025). Predicting student performance and enhancing learning outcomes: a data-driven approach using educational data mining techniques. *Computers*, 14(3), 83. DOI: <https://doi.org/10.3390/computers14030083>
- [17] Maraza-Quispe, B., Damian Valderrama-Chauca, E., Henry Cari-Mogrovejo, L., & Milton Apaza-Huanca, J. (2021, October). Predictive model of student academic performance from LMS data based on learning analytics. In *Proceedings of the 13th International Conference on Education Technology and Computers* (pp. 13-19). DOI: <https://doi.org/10.1145/3498765.3498768>
- [18] Al-Obaidi, H. (2023). Student Grade Prediction using Machine Learning Methods. <https://repository.rit.edu/theses/12213>
- [19] Qazdar, A., Qassimi, S., HASSIDI, O., Hafidi, M., Abdelwahed, E. H., & Melk, Y. (2022). Learning analytics for tracking student progress in lms. DOI: <https://doi.org/10.21203/rs.3.rs-1505417/v1>
- [20] Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human behavior*, 104, 106189. DOI: <https://doi.org/10.1016/j.chb.2019.106189>
- [21] Alyahyan, E., & Düşteğör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International journal of educational technology in higher education*, 17(1), 3. DOI: <https://doi.org/10.1186/s41239-020-0177-7>
- [22] Patel, K. K., & Amin, K. (2024). Predictive modeling of dropout in MOOCs using machine learning techniques. *The Scientific Temper*, 15(02), 2199-2206. DOI: <https://doi.org/10.58414/SCIENTIFICTEMPER.2024.15.2.32>
- [23] Maloney, S., Axelsen, M., Galligan, L., Turner, J., Redmond, P., Brown, A., ... & Lawrence, J. (2022). Using LMS Log Data to Explore Student Engagement with Coursework Videos. *Online Learning*, 26(4), 399-423.
- [24] Brahim, G. B. (2022). Predicting student performance from online engagement activities using novel statistical features. *Arabian Journal for Science and Engineering*, 47(8), 10225-10243. DOI: <https://doi.org/10.1007/s13369-021-06548-w>
- [25] Alnasyan, B., Basher, M., & Alassafi, M. (2024). The power of Deep Learning techniques for predicting student performance in Virtual Learning Environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6, 100231. DOI: <https://doi.org/10.1016/j.caeai.2024.100231>