

Artificial Intelligence-Based Prediction of Learning Difficulties Through Educational Data Analytics

Elena Rinaldi

Research Fellow, European Institute for Adaptive Learning Research, Italy

Abstract

Educational institutions increasingly generate digital information through learning management systems, online assessments, attendance platforms, assignment submissions, and student interaction records. These data can contain early indicators of academic difficulty before poor final outcomes become apparent. Artificial intelligence and educational data analytics provide a methodological basis for combining multiple indicators and estimating which learners may require additional instructional support. Such prediction should be interpreted as an early-support mechanism rather than as a diagnostic classification of students. This hypothetical methodological study examines artificial intelligence-based prediction of learning difficulties through educational data analytics. A fictional dataset containing 1,800 student records and twenty-two educational variables is divided into training, validation, and independent testing subsets. Three conceptual predictive models are compared: logistic regression, support vector machine, and random forest classification. Predictor variables include assessment performance, assignment completion, learning-management-system activity, attendance consistency, response latency, repeated-error patterns, and recent changes in engagement.

The illustrative analysis evaluates accuracy, precision, recall, and F1 score using a hypothetical test set containing ninety students designated as requiring additional academic support and 210 students without the defined support-risk label. Under the assigned values, the random forest model produces the strongest overall performance. A second analysis examines prediction quality at different stages of a fictional academic term and indicates that model reliability improves as additional educational evidence becomes available. The study concludes that AI-based educational analytics can support earlier recognition of learners who may benefit from instructional assistance. However, predictions should supplement teacher judgment rather than determine student capability, placement, diagnosis, or opportunity. Reliable use requires transparent labeling, representative data, privacy protection, subgroup monitoring, human review, and educationally meaningful intervention procedures. All numerical findings reported in this manuscript are illustrative rather than empirically observed.

Keywords: Artificial Intelligence, Educational Data Analytics, Learning Difficulties, Educational Data Mining, Learning Analytics, Student Performance Prediction, Early Warning Systems, Machine Learning, Academic Support, Predictive Analytics.

1. Introduction

Digital learning environments generate substantial information concerning how students interact with educational materials, complete assessments, participate in learning activities, and progress through courses. These records can provide indicators of changes in academic performance before final grades are available. Educational data mining and learning analytics have therefore developed as interdisciplinary areas concerned with extracting meaningful patterns from educational information. Baker and Yacef described prediction as an increasingly important educational data-mining task, while Romero and Ventura documented the wider development of computational methods for educational analysis.

The prediction of academic difficulty is especially relevant when institutions seek to provide timely assistance rather than waiting until a learner has already failed a course. Kotsiantis et al. demonstrated that machine learning could be used to predict performance within distance-learning environments. Lykourantzou et al. later investigated machine-learning techniques for identifying students at risk of withdrawing from e-learning courses. These studies established that behavioral, assessment, and participation information can contribute to predictive educational models when outcomes are defined carefully.

Learning-management-system activity provides another important source of educational evidence. Macfadyen and Dawson demonstrated that LMS interaction variables could support an early-warning model for identifying students at risk of unsuccessful course outcomes. Arnold and Pistilli subsequently described an academic-analytics initiative in which predictive information was used to provide students and instructors with early performance signals. Such systems illustrate a transition from retrospective educational reporting toward proactive support-oriented analytics.

Artificial intelligence can integrate multiple indicators simultaneously, but prediction must not be confused with diagnosis. Academic difficulty may emerge from temporary misunderstanding, insufficient prerequisite knowledge, irregular participation, assessment design, language barriers, external circumstances, or other educational factors. A statistical risk score cannot determine whether a learner has a developmental or clinical learning condition. AI-based systems should therefore identify patterns requiring pedagogical attention rather than assign fixed labels concerning student ability or disability.

The present hypothetical study examines three machine-learning approaches for predicting educational support needs from longitudinal academic data. Logistic regression provides an interpretable baseline, support vector machine represents nonlinear classification, and random forest models interactions among multiple educational indicators. Performance is evaluated both overall and at different points during an academic term. No actual students were analyzed, and the entire dataset and numerical results are constructed solely to demonstrate an academically defensible methodological framework.

2. Objectives of the Study

The principal objective of this study is to examine whether artificial intelligence-based educational data analytics can identify early patterns associated with emerging academic learning difficulties. The investigation focuses on predictive indicators, comparative model performance, timing of prediction, and responsible educational interpretation.

2.1 Identification of Educational Risk Indicators

The first objective is to examine educational variables that may signal an increased need for learning support. The conceptual feature set includes declining assessment scores, repeated errors across related concepts, delayed assignment submission, incomplete learning activities, reduced LMS engagement, irregular attendance, low practice frequency, and changes in response time. These indicators are intended to represent observable learning behavior rather than permanent characteristics of the learner.

Educational indicators should be interpreted contextually because an isolated low score or temporary reduction in participation does not establish persistent difficulty. The methodology therefore combines several observations across time rather than relying on one event. A student may miss an assignment for reasons unrelated to learning, while repeated difficulties across assessment, participation, and practice activities may provide stronger evidence that targeted academic support should be considered.

2.2 Comparison of Artificial Intelligence Models

The second objective is to compare logistic regression, support vector machine, and random forest classification under identical hypothetical training and testing conditions. Logistic regression provides a transparent reference model in which predictor effects can be examined directly. Support vector machine provides nonlinear decision boundaries, while random forest can model interactions among multiple educational features without requiring a predefined linear relationship.

Model comparison is based on accuracy, precision, recall, and F1 score. Recall is particularly important because failing to identify a learner who genuinely requires additional support may delay intervention. Precision is equally relevant because excessive false-positive identification can unnecessarily label students and increase educator workload. The study therefore avoids treating overall accuracy as the only criterion for evaluating educational prediction.

2.3 Evaluation of Prediction Timing and Educational Use

The third objective is to examine how prediction performance changes as additional course information becomes available. Very early prediction may provide more time for intervention but may rely on limited educational evidence. Predictions generated later in the term can use richer behavioral and assessment histories, although the remaining time for effective instructional support becomes shorter. The timing of prediction therefore creates an important educational trade-off.

The study also aims to clarify how predictions should be used. Risk scores are intended to trigger review by educators, not automatic academic penalties or irreversible decisions. Teachers should examine recent work, learning context, and student needs before determining whether additional tutoring, feedback, study guidance, or instructional adaptation is appropriate. AI should support educational decision-making rather than replace professional assessment.

3. Review of Literature

3.1 Educational Data Mining and Learning Analytics

Baker and Yacef reviewed the development of educational data mining and identified prediction, clustering, relationship mining, and discovery with models as important research directions. Romero and Ventura provided a broader review of educational data mining and documented the application of classification, association, clustering, and statistical techniques to educational environments. These

works established a conceptual foundation for using computational methods to understand student behavior and learning outcomes.

Learning analytics developed alongside educational data mining with a stronger emphasis on institutional decision support and learning processes. Campbell et al. described academic analytics as the use of large institutional datasets and predictive modeling to generate actionable information. Siemens and Baker later discussed the relationship between learning analytics and educational data mining, emphasizing opportunities for collaboration between the two research communities. Ferguson similarly reviewed the development and emerging challenges of learning analytics.

3.2 Prediction of Academic Risk and Performance

Kotsiantis et al. compared machine-learning techniques for predicting performance among distance-learning students and demonstrated that educational variables could support early identification of students likely to experience academic difficulty. Dekker et al. subsequently examined student dropout prediction and showed the usefulness of interpretable classifiers and cost-sensitive analysis. Lykourantzou et al. used multiple machine-learning methods and combined classifiers to identify dropout risk within e-learning environments.

Márquez-Vera et al. investigated prediction of school failure and dropout using data-mining techniques and highlighted the importance of class balancing, attribute selection, and interpretable classification. These studies demonstrate that educational prediction is affected not only by algorithm choice but also by data quality, class distribution, feature representation, and the practical cost associated with different prediction errors.

3.3 Early-Warning Systems and Educational Intervention

Macfadyen and Dawson demonstrated that LMS-generated activity data could support an early-warning model for identifying students at risk of failure. Their work emphasized behavioral indicators such as assessment completion and online participation. Arnold and Pistilli later described Course Signals, an academic-analytics system designed to communicate early performance information to students and instructors. These initiatives illustrate how predictive analytics can be connected to educational action rather than functioning only as retrospective analysis.

Jayaprakash et al. examined an open-source early-alert initiative involving prediction of academic risk and intervention. Their work emphasized the scalability, portability, and practical implementation of analytics across educational institutions. Collectively, the literature suggests that prediction becomes educationally meaningful only when it is linked to appropriate support processes. Accurate identification without timely and constructive intervention provides limited direct benefit to learners.

4. Research Methodology

4.1 Research Design and Predictive Framework

The present study adopts a hypothetical binary-classification design. The positive class represents students who meet a predefined educational support-risk criterion based on persistent low academic performance combined with multiple indicators of learning difficulty. The negative class represents students who do not meet this threshold during the observation period. The classification is explicitly educational and does not represent disability, psychological condition, or clinical diagnosis.

Three conceptual machine-learning models are compared. Logistic regression uses standardized predictor variables with regularization. The support vector machine uses a radial-basis-function kernel to model nonlinear relationships. Random forest combines multiple decision trees constructed from bootstrapped training observations and randomized subsets of predictors. Hyperparameters are assumed to be selected using only the validation partition, with the test set reserved for final evaluation.

4.2 Hypothetical Student Dataset and Variables

The fictional dataset contains 1,800 anonymized student records and twenty-two educational variables. One thousand two hundred records are assigned to training, three hundred to validation, and three hundred to independent testing. The test partition contains ninety records labeled as requiring additional academic support and 210 records not meeting the hypothetical risk criterion. No record represents a real student, institution, or learning platform.

Variables include recent assessment average, assessment-score trend, assignment completion rate, late-submission frequency, quiz retry count, concept-level error recurrence, attendance consistency, LMS login frequency, study-session frequency, resource-view count, response latency, practice completion, and recent changes in engagement. Direct identity variables and sensitive diagnostic information are excluded. A genuine study would require appropriate consent, governance, access control, and validation of every educational variable before use.

4.3 Evaluation Procedure and Ethical Interpretation

The same 300-record test set is used for all three models. True positives represent students correctly identified as meeting the educational support-risk criterion, while false negatives represent support-risk students not identified by the model. False positives represent students incorrectly flagged for additional review. Accuracy, precision, recall, and F1 score are calculated from the resulting hypothetical confusion matrices.

A second analysis evaluates the random-forest configuration at Weeks 2, 4, 6, and 8 of the fictional academic term. Only information available up to each prediction point is included. This prevents later assessment outcomes from leaking into earlier predictions. The model output is conceptualized as a support-review signal; educators would still be required to inspect the student's work and context before initiating any instructional response.

5. Analysis

5.1 Comparative Prediction Performance

The first hypothetical analysis compares the three machine-learning models on the 300-record test set. Logistic regression correctly identifies sixty-five of the ninety support-risk students, while the support vector machine identifies seventy-two. Random forest identifies seventy-eight and generates twenty-one false-positive classifications. These assigned values illustrate how nonlinear interactions among engagement, assessment, and participation variables may improve prediction when learning difficulties are expressed through several simultaneous patterns.

The random forest records hypothetical accuracy of 89.00%, precision of 78.79%, recall of 86.67%, and F1 score of 82.54%. The support vector machine produces an F1 score of 77.01%, while logistic regression reaches 70.27%. These differences are illustrative and do not establish that random

forests are universally superior in educational prediction. Actual results would depend on institutional context, label definition, course design, student population, and feature quality.

Table 1: Hypothetical Prediction of Educational Learning-Support Risk

Predictive Model	TP	FP	FN	TN	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic Regression	65	30	25	180	81.67	68.42	72.22	70.27
Support Vector Machine	72	25	18	185	85.67	74.23	80.00	77.01
Random Forest	78	21	12	189	89.00	78.79	86.67	82.54

Note: All student records, classifications, and performance values are fictional. The positive class indicates a need for academic support review, not a diagnosis of learning disability.

5.2 Early Prediction Across the Academic Term

The second hypothetical analysis evaluates random-forest performance using only information available at different stages of the academic term. At Week 2, the model receives relatively limited evidence and records an illustrative F1 score of 63.94%. Performance increases at Weeks 4 and 6 as additional assessments and participation records become available. By Week 8, the F1 score reaches 82.54%, corresponding to the complete hypothetical feature window used in the primary comparison.

The results illustrate a central early-warning trade-off. Earlier predictions provide more time for support but may produce greater uncertainty because relatively little academic evidence has accumulated. Later predictions can be more reliable but reduce the period available for instructional response. Educational institutions should therefore select prediction timing according to both statistical reliability and the practical ability to provide meaningful intervention.

Table 2: Hypothetical Random-Forest Performance by Prediction Stage

Prediction Stage	Precision (%)	Recall (%)	F1 Score (%)
Week 2	66.00	62.00	63.94
Week 4	72.00	76.00	73.95
Week 6	76.50	81.10	78.73
Week 8	78.79	86.67	82.54

Note: Prediction stages and all numerical performance values are illustrative rather than empirically measured.

6. Discussion

6.1 Educational Data Patterns and Learning Support

The hypothetical findings illustrate how multiple educational indicators can provide a more informative view of learning difficulty than one assessment score alone. A student experiencing repeated concept-level errors, declining assessment performance, incomplete practice, and reduced participation may warrant instructional review even before final course outcomes are known. Educational data mining enables these patterns to be examined systematically across large student populations.

However, prediction should remain sensitive to educational context. Reduced LMS activity may reflect efficient offline study rather than disengagement, while repeated attempts can represent productive persistence rather than difficulty. Predictive variables should therefore be interpreted as statistical signals rather than direct measures of ability.

6.2 Comparative Model Performance and Interpretability

The assigned random-forest performance illustrates the potential value of nonlinear models when relationships among educational variables are complex. Assessment decline may become important only when combined with reduced practice or repeated conceptual errors. Tree-based ensembles can represent such interactions more flexibly than a strictly linear model. Nevertheless, predictive improvement should not automatically justify greater algorithmic complexity.

Educational systems require interpretability because instructors need to understand why a student has been flagged. Logistic regression provides relatively direct coefficient-based interpretation, whereas random forests require additional explanatory analysis. Variable-importance measures or local explanation methods can help, but they must be interpreted carefully. An educational explanation should identify actionable indicators—such as missing practice or recurring errors—rather than merely report a numerical risk probability.

6.3 Early Warning and Human-Centered Intervention

The timing analysis reinforces findings from early-warning literature showing that predictive information is useful primarily when it reaches educators early enough to support action. Macfadyen and Dawson demonstrated the potential of LMS information for early identification, while Arnold and Pistilli and Jayaprakash et al. emphasized intervention-oriented analytics. Predictive accuracy alone does not guarantee student benefit because the institutional response determines whether an alert produces meaningful support.

A responsible response may include additional feedback, tutoring, revision materials, clarification of prerequisite concepts, study-planning assistance, or a conversation with the learner. The intervention should not assume that the student is incapable of succeeding. Educational analytics should therefore be designed around supportive escalation rather than punitive classification. The learner should remain an active participant in any decisions arising from predictive information.

7. Challenges

7.1 Label Definition and Measurement Validity

A major challenge is defining what constitutes a “learning difficulty” within educational data. Low grades alone may reflect missed work, unfamiliar assessment formats, inconsistent attendance, or temporary circumstances rather than persistent conceptual difficulty. If the target label is poorly defined,

machine learning can produce statistically consistent predictions of an educational construct that lacks validity. The quality of the target therefore matters as much as classifier performance.

A stronger approach requires combining several educational indicators and reviewing labels with subject educators. The present hypothetical design uses persistent low performance together with recurring evidence across multiple learning behaviors. Even this definition remains an academic-support construct rather than a diagnosis. Real research should document the exact labeling procedure and distinguish course difficulty, disengagement, language-related challenges, and other sources of unsuccessful performance whenever the available data permit.

7.2 Data Bias, Privacy, and Student Fairness

Educational datasets can reflect differences in access to technology, prior preparation, course design, and opportunities for participation. Models may therefore assign different risk probabilities to students for reasons unrelated to their underlying capacity to learn. Predictive systems should be evaluated across relevant student groups and educational settings to identify systematic error differences. A high overall accuracy can conceal weaker performance for smaller or underrepresented populations.

Privacy is equally important because educational records can contain detailed behavioral information. Data should be collected only for legitimate educational purposes and protected through appropriate technical and administrative controls. Predictive outputs should be accessible only to personnel with a valid educational role. Risk scores should not become permanent student labels or be reused for unrelated decisions without appropriate justification and governance.

7.3 Generalization and Over-Reliance on Automation

A predictive model developed within one course may not transfer reliably to another course because assessment structures, learning activities, content difficulty, and participation expectations differ. Even within the same institution, an indicator such as discussion-board activity may be highly relevant in one course and irrelevant in another. External and cross-course validation are therefore necessary before a prediction model is applied broadly.

Over-reliance on automated prediction creates an additional challenge. Teachers may give excessive weight to model scores even when contextual information suggests a different interpretation. The system should communicate uncertainty and provide the educational evidence contributing to the flag. AI-based early warning is most defensible when it broadens educator awareness rather than replacing observation, dialogue, and professional judgment.

8. Conclusion

Artificial intelligence-based educational data analytics provides a useful methodological framework for identifying early patterns associated with academic learning-support needs. The present hypothetical study compared logistic regression, support vector machine, and random forest models using assessment, participation, engagement, and learning-behavior indicators. The objective was to support earlier pedagogical review rather than classify students permanently or diagnose learning disabilities.

The illustrative analysis assigned the strongest overall performance to the random-forest model and demonstrated that predictive reliability increased as additional educational evidence became available during the academic term. Every numerical value reported in the analysis is hypothetical and

should not be interpreted as empirical evidence. Actual performance claims would require real student data, validated target definitions, temporal testing, subgroup analysis, and independent replication.

The study concludes that AI can contribute to early academic support when predictive models are transparent, educationally grounded, privacy-conscious, and integrated with human judgment. Future empirical research should emphasize construct validity, fairness across student populations, cross-course generalization, interpretable prediction, and evaluation of the interventions triggered by risk alerts. The ultimate criterion should be whether educational analytics helps learners receive appropriate and timely support.

References

1. Agrusti, F., Bonavolontà, G., & Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. *Journal of E-Learning and Knowledge Society*, 15(3), 161–182. <https://doi.org/10.20368/1971-8829/1135017>
2. 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, Article 3. <https://doi.org/10.1186/s41239-020-0177-7>
3. Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson and B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4
4. Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In *Proceedings of the 5th Annual Future Business Technology Conference* (pp. 5–12). EUROSIS.
5. Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59, 64–71. <https://doi.org/10.1007/s11528-014-0822-x>
6. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: A systematic literature review. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education* (pp. 175–199). Association for Computing Machinery. <https://doi.org/10.1145/3197091.3197097>
7. Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, 36, 469–478. <https://doi.org/10.1016/j.chb.2014.04.002>
8. Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18(5), 411–426. <https://doi.org/10.1080/08839510490442058>
9. Kuzilek, J., Hlostá, M., & Zdrahal, Z. (2017). Open University Learning Analytics dataset. *Scientific Data*, 4, Article 170171. <https://doi.org/10.1038/sdata.2017.171>
10. Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1909–1918). Association for Computing Machinery. <https://doi.org/10.1145/2783258.2788620>

11. Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Fardoun, H. M., & Ventura, S. (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>
12. Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, 17(4), 49–64.
13. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), Article e1355. <https://doi.org/10.1002/widm.1355>
14. Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10), 1380–1400. <https://doi.org/10.1177/0002764213498851>
15. Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414–422. <https://doi.org/10.1016/j.procs.2015.12.157>
16. Xiao, W., Ji, P., & Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. *Engineering Reports*, 4(5), Article e12482. <https://doi.org/10.1002/eng2.12482>
17. Zacharis, N. Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education*, 27, 44–53. <https://doi.org/10.1016/j.iheduc.2015.05.002>
18. Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, 47, 157–167. <https://doi.org/10.1016/j.chb.2014.05.038>
19. Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98–110. <https://doi.org/10.1016/j.chb.2018.07.027>
20. Lu, Y., Yeom, S., Maktoubian, J., Rahman, M. M., & Kim, S.-H. (2025). Improve student risk prediction with clustering techniques: A systematic review in education data mining. *Education Sciences*, 15(12), Article 1695. <https://doi.org/10.3390/educsci15121695>