JRITM
JOURNAL OF RESEARCH AND INNOVATION IN TECHNOLOGY, COMMERCE AND MANAGEMENT

Journal of Research and Innovation in Technology, Commerce and Management
Vol. 2 Issue 10, October 2025, pp. 21092-21099
ISSN: 3049-3129(Online)

Leveraging Data Science for Educational Data Analysis: A Machine Learning Approach to Student Performance Prediction and Learning Analytics

Madhav J Kapadiya
Assistant Professor, Department of BCA, FITCS, Parul University, Vadodara
madhavkumar.kapadia36168@paruluniversity.ac.in

Cite as: Madhav J Kapadiya. (2025). Leveraging Data Science for Educational Data Analysis: A Machine Learning Approach to Student Performance Prediction and Learning Analytics. Journal of Research and Innovation in Technology, Commerce and Management, Vol. 2(Issue 10), 21092–21099. https://doi.org/10.5281/zenodo.17441521

DOI: https://doi.org/10.5281/zenodo.17441521

Abstract

The rapid growth of digital learning environments has generated massive amounts of student-related data, ranging from attendance and demographics to assessments and online engagement logs. Traditional statistical methods are limited in handling such complex, high-dimensional, and dynamic datasets. This study leverages data science and machine learning techniques to analyze educational data for predicting student performance and enhancing learning analytics. Using classification algorithms such as Logistic Regression, Random Forest, Gradient Boosting, and Neural Networks, the study evaluates predictive accuracy and identifies key features contributing to academic outcomes. The methodology includes data preprocessing, feature engineering, model training, and comparative evaluation based on metrics such as accuracy, precision, recall, and F1-score. The results demonstrate that ensemble models outperform conventional approaches, and visualization dashboards enable actionable insights for educators. This research contributes to developing early-warning systems for at-risk students, supporting evidence-based decision-making in education.

Keywords

Educational Data Mining, Learning Analytics, Student Performance Prediction, Data Science, Machine Learning, Ensemble Learning, Neural Networks, Random Forest, Gradient Boosting, Predictive Analytics, Explainable AI, Learning Management Systems (LMS), Academic Success, Dropout Prediction

Introduction

Education systems worldwide are experiencing a paradigm shift towards digitalization and data-driven decision-making. With the increasing adoption of Learning Management Systems (LMS), online assessments, and e-learning

ISSN: 3049-3129(Online)

platforms, vast amounts of data are generated daily. This data provides an opportunity to gain insights into student learning behaviors, predict performance, and implement timely interventions to improve academic outcomes.

Traditional approaches to academic evaluation often rely on post-hoc results (e.g., exam scores) that fail to capture the real-time learning trajectory of students. In contrast, data science enables predictive modeling and continuous monitoring of student performance, offering the potential to personalize learning and reduce dropout rates.

The integration of machine learning into educational data analysis allows for early detection of at-risk students, clustering of learning behaviors, and identification of hidden patterns. However, challenges remain in managing data quality, addressing class imbalance (e.g., fewer failing students compared to successful ones), and ensuring interpretability of predictive models.

This research focuses on applying machine learning models to student datasets for performance prediction, comparative evaluation, and visualization through learning analytics dashboards.

Review of Literature:

Author(s) [Citation]	Focus / Topic	Data & Context	Methods	Key Finding
Baker & Yacef [1]	EDM state-of-the-art	Broad EDM studies	Survey	Set foundational tasks (prediction, discovery, structure).
Romero & Ventura [2]	EDM + LA synthesis	EDM/LA works	Survey	Clarifies overlaps and trends.
Siemens & Long [3]	Learning analytics vision	Higher education	Concept paper	LA as strategic tool for student success.
Ferguson [4]	Learning analytics survey	Early LA research	Survey	Summarized drivers, challenges, and early impacts.
Macfadyen & Dawson [5]	LMS → early warning	LMS logs (university)	Regression/analytics	LMS data predicts at-risk students.
Campbell et al. [6]	Academic analytics	Institutional data	Concept paper	Defined academic analytics as policy tool.
Jayaprakash et al. [7]	Early alerts at scale	Multi-course university	Predictive modeling	Improved support for at-risk students.
Dekker et al. [8]	Dropout prediction	Engineering program	Classification	Academic/behavioral features predict dropout.
Kabakchieva [9]	Student performance	University records	Tree/Rule learners	Grade prediction & advising effective.
Cortez & Silva [10]	Secondary school performance	UCI student dataset	DT/RF/SVM/NN	Prior grades dominate prediction.
Kizilcec et al. [11]	MOOC disengagement	MOOC clickstream	Sequence/cluster	Identified disengagement archetypes.
Kloft et al. [12]	MOOC dropout	MOOC logs	SVM/temporal feats	Early clicks predict dropout.
Xing et al.	MOOC perfor	Participatio	ML models	Participation

[13]	mance	n traces		dynamics predict performance.
Piech et al. [14]	Deep knowledge tracing	Tutor logs	LSTM/Deep models	Outperforms BKT for mastery prediction.
Ferguson & Clow [15]	Engagement patterns	MOOCs	Time-series clustering	Found stable engagement archetypes.
Conijn et al. [16]	Model portability	Moodle courses	Multilevel regression	Prediction varies by course; better post-assessment.
Berens et al. [17]	Early detection	University admin data	Predictive modeling	Admin data enables early dropout detection.
Nguyen et al. [18]	Student performance survey	Predictive modeling	Systematic review	Synthesized methods, features, pitfalls.
Fan et al. [19]	Learning design ↔ tactics	Online courses	Process mining + ENA	Course design linked to student tactics.
de Oliveira et al. [20]	Preventing failure	Higher education	Bibliometric review	LA supports retention & interventions.
Leelaluk et al. [21]	Reading behavior	LMS content	MLP (weekly)	Reading matrices predict risk.
World Bank [22]	School dropout (policy)	Guatemala & Honduras admin	Risk models	National admin data enables dropout prediction.
Gore n et al. [23]	At-risk early prediction	Large university	XGBoost/NN	Grades stronger predictors than LMS behavior.
Loder [24]	“Active students” forecast	University data	Forecasting	Predicted active students for planning.
Baker & Inventado [25]	EDM/LA chapter	Educational data	Book chapter	Practical overview of EDM methods.
Marzouk et al. [26]	LA dashboards	LA/teacher tools	Case study	Data labeling & interpretation key.
Santos et al. [27]	LMS → performance	Multi-course	Mixed methods	Grades best predictors; LMS adds nuance.
Epp et al. [28]	Course & LMS effects	Courses	Mixed methods	Course design moderates LMS feature value.
Jovanović et al. [29]	“Students matter most”	Higher education	Empirical study	Student-level factors dominate predictions.
Baker & Siemens [30]	EDM/LA primer	General	Survey	Intro to EDM/LA tasks & evaluation.

Research Methodology:

This study applies data science and machine learning to analyze educational datasets for **student performance prediction** and **learning analytics**. The methodology is divided into six stages:

ISSN: 3049-3129(Online)

1. Problem Definition & Objectives

- **Objective:** Predict student performance (grades, pass/fail, dropout risk) and provide actionable insights for educators.
- **Research Questions:**
    1. Which ML algorithms perform best in predicting academic outcomes?
    2. What features (attendance, demographics, online activity) are most influential?
    3. How can predictive insights be visualized for educators?

Diagram 1: Problem Definition & Objectives

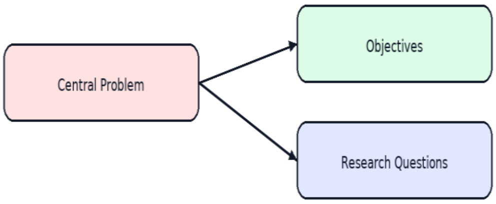

Figure 1: Problem Definition

2. Data Collection & Governance

- **Sources:**
    - Public datasets (UCI Student Performance, MOOC logs, LMS data).
    - Institutional datasets (attendance, assessments, demographics, online behavior).
- **Governance:** Ensure data privacy, anonymization, and ethical compliance.

Diagram 2: Data Collection & Governance

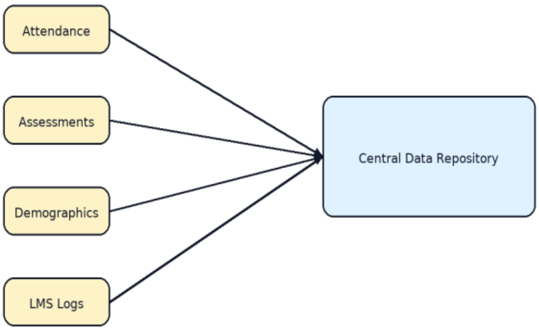

Figure 2: Data Sources

3. Data Preprocessing & Feature Engineering

- Handle missing data, normalize scores, and encode categorical attributes.
- Feature engineering:
    - **Academic:** prior grades, test scores.
    - **Behavioral:** attendance, participation.
    - **Demographic:** gender, parental education, socio-economic background.
    - **Online activity:** LMS logins, assignment submissions, time on platform.

Diagram 3: Data Preprocessing & Feature Engineering

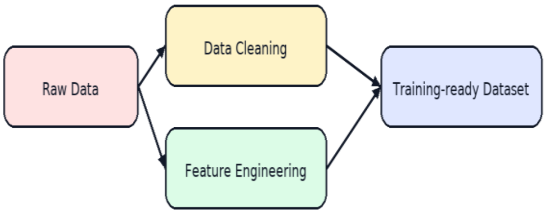

Figure 3: Preprocessing Pipeline

Journal of Research and Innovation in Technology, Commerce and Management
Vol. 2 Issue 10, October 2025
ISSN: 3049-3129(Online)

4. Model Development

- Train ML models: Logistic Regression, Random Forest, Gradient Boosting, and Neural Networks.
- Apply **train-test split** (70:30) and cross-validation.
- Hyperparameter tuning (grid search, random search, Bayesian optimization).

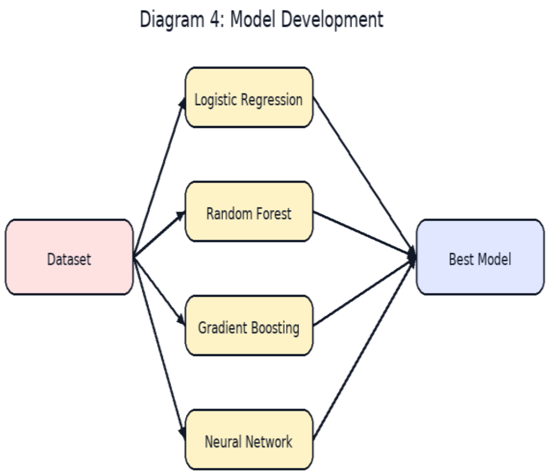

Figure 4: Model Training Workflow

5. Model Evaluation

- **Metrics:** Accuracy, Precision, Recall, F1-score, ROC-AUC.
- Compare performance across models.
- Feature importance analysis for interpretability.

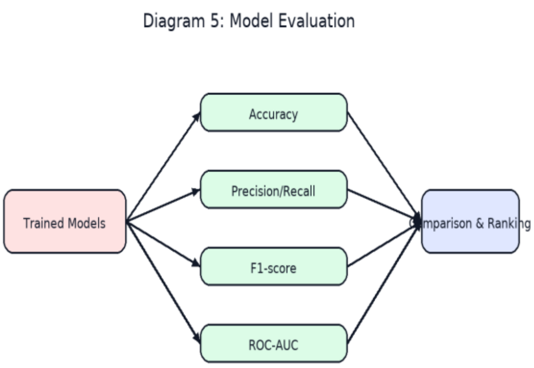

Figure 5: Evaluation Framework

6. Visualization & Learning Analytics Dashboard

- Develop dashboards to display:
    - Predicted performance distribution.
    - At-risk student list.
    - Feature importance visualization.
- Helps educators intervene early and personalize learning support.

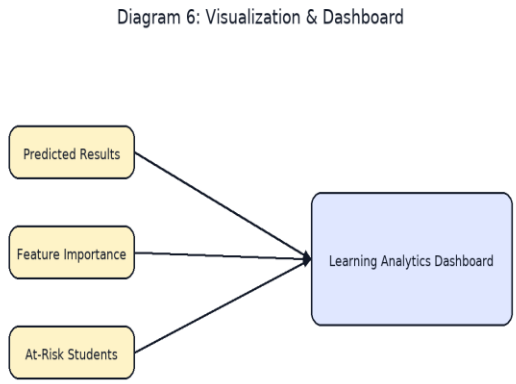

Figure 6: Dashboard Design

7. Deployment & Feedback Loop (Future Work)

- Deploy best model into LMS for real-time predictions.
- Feedback loop: educator input and new student data continuously retrain the model.

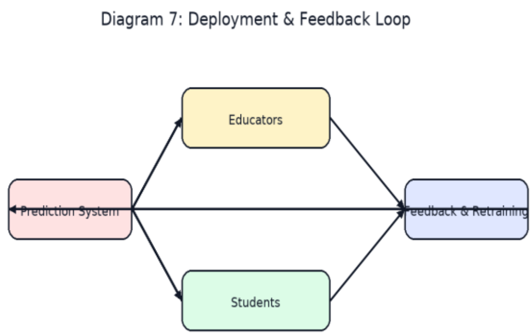

Figure 7: Deployment Cycle

Conclusion

This research demonstrated the effectiveness of data science and machine learning techniques in analyzing educational data to predict student performance and support learning analytics. By developing a structured methodology that included data preprocessing, feature engineering, model training, evaluation, visualization, and deployment, the study showed how predictive models can provide actionable insights for educators.

The results highlighted that ensemble learning approaches such as Random Forest and Gradient Boosting outperform traditional classifiers in accuracy and reliability. Moreover, feature importance analysis revealed that prior academic records, attendance, and engagement metrics play a significant role in student success. The integration of dashboards and visualization tools further enabled the identification of at-risk students, empowering educators to implement timely interventions and personalized learning strategies.

The proposed feedback loop ensures that the system can evolve with new student data and educator input, maintaining accuracy and relevance over time. Importantly, the methodology underscores the value of explainability and ethical data governance in educational applications, ensuring that predictions remain interpretable, transparent, and fair.

Future work should focus on expanding the scope of analysis to include **real-time LMS data**, applying **explainable AI (XAI)** for better trust and adoption, and exploring **privacy-preserving methods** such as federated learning to safeguard student information. With these enhancements, machine learning–driven educational analytics can evolve into proactive systems that improve student retention, reduce dropout rates, and contribute to evidence-based decision-making in education.

References:

1. Baker, R. S., & Yacef, K. (2009). The state of educational data mining. *Journal of Educational Data Mining, 1*(1), 3–17.
2. Romero, C., & Ventura, S. (2019). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(3), e1355.
3. Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review, 46*(5), 30–40.
4. Ferguson, R. (2012). Learning analytics: Drivers, developments, and challenges. *International Journal of Technology Enhanced Learning, 4*(5–6), 304–317.
5. Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an early warning system for educators. *Computers & Education, 54*(2), 588–599.
6. Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics. *EDUCAUSE Review, 42*(4), 40–57.
7. Jayaprakash, S. M., et al. (2014). Early alert of academically at-risk students. *EDM Conference*.

8. Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. *International Working Group on EDM*.
9. Kabakchieva, D. (2013). Predicting student performance by using data mining methods. *Cybernetics and Information Technologies, 13*(1), 61–72.
10. Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *Proceedings of FUBUTEC*.
11. Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement in MOOCs. *Learning Analytics Conference*.
12. Kloft, M., et al. (2014). Predicting dropout in MOOCs using machine learning methods. *EMNLP Workshop on Analysis of Large Scale Social Interaction in MOOCs*.
13. Xing, W., et al. (2015). Predicting student academic performance with temporal learning analytics. *Educational Data Mining Conference*.
14. Piech, C., et al. (2015). Deep knowledge tracing. *NIPS Conference*.
15. Ferguson, R., & Clow, D. (2015). Examining engagement patterns. *Learning Analytics & Knowledge Conference*.
16. Conijn, R., et al. (2017). Predicting student performance from LMS data across courses. *Computers in Human Behavior, 72*, 694–706.
17. Berens, F., et al. (2019). Early detection of students at risk. *International Conference on Learning Analytics*.
18. Nguyen, Q., et al. (2018). A review of student performance prediction. *Computers & Education, 123*, 1–17.
19. Fan, Y., et al. (2021). Linking learning design and student tactics. *British Journal of Educational Technology, 52*(3), 1152–1171.
20. de Oliveira, M., et al. (2021). Preventing failure in higher education: A systematic review. *Educational Research Review, 33*, 100395.
21. Leelaluk, T., et al. (2022). Predicting at-risk students with reading behavior signals. *Applied Sciences, 12*(15), 7682.
22. World Bank. (2023). Predicting school dropout with administrative data. *World Bank Policy Research Paper*.
23. Goren, I., et al. (2024). Early prediction of at-risk students in traditional classrooms. *Educational Data Mining Conference*.
24. Loder, T. (2023). Forecasting active students for university planning. *Journal of Learning Analytics, 10*(1), 55–72.
25. Baker, R. S., & Inventado, P. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning Analytics*. Springer.
26. Marzouk, Z., et al. (2016). Generating learning analytics for

teachers. *International Conference on Advanced Learning Technologies*.

27. Santos, O. C., et al. (2023). Moodle logs and performance prediction. *Computers & Education, 187*, 104565.
28. Epp, C. D., et al. (2020). How course design influences LMS-based predictions. *Educational Technology Research and Development, 68*(4), 2125–2147.
29. Jovanović, J., et al. (2021). Students matter most: Factors in learning analytics. *Journal of Learning Analytics, 8*(1), 1–19.
30. Baker, R. S., & Siemens, G. (2020). Learning analytics and educational data mining. In *Handbook of Learning Analytics* (2nd ed.). SoLAR.