International Journal of Information System & Innovative Technology 4(1) (2025) 11-14

IJISIT

http:ejournal.gevivapublisher.orgindex.phpijisit

Predicting Student Academic Performance Using Random Forest Regression: A Case Study on LMS Behavioral Data

[1]Rijois Iboy Erwin Saragih, Universitas Methodist Indonesia, Indonesia
[2]Thandy Simanjuntak, Boston University, USA
[3]Joe Silitonga, Ericsson Telecomunication Pte Ltd, Singapore
Correspondence: E-mail: rijoissaragih@gmail.com

Article Info

Article history:
Received May 13, 2025
Revised June 14, 2025
Accepted June 20, 2025

Keywords:
Student performance,
Machine learning,
Random forest,
Regression,
LMS behavior

ABSTRACT

In the evolving landscape of digital education, Learning Management Systems (LMS) have become pivotal in managing student engagement and academic resources. These platforms not only facilitate course delivery but also log extensive behavioral data, including attendance rates, quiz performances, LMS usage time, and forum activities. Leveraging this data, educators and institutions can enhance academic outcomes through predictive analytics. This study investigates the use of Random Forest Regression, a machine learning technique, to predict student final grades based on LMS behavioral data. A synthetic dataset comprising 100 student records was used, each containing features that reflect engagement and performance. The data underwent standard preprocessing procedures including normalization and partitioning into training and testing sets. The Random Forest model was trained and evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as performance metrics. The model achieved a MAE of 4.57 and RMSE of 5.90, indicating a high level of predictive accuracy. Feature importance analysis revealed that average quiz score and attendance rate were the most significant predictors, followed by LMS time and forum activity. These findings demonstrate the effectiveness of ensemble learning methods in educational settings and support the integration of predictive systems in LMS platforms for real-time academic monitoring. Such systems could provide early alerts for at-risk students and assist educators in designing targeted interventions. This research contributes to the field of Educational Data Mining by validating the practical utility of Random Forest Regression in supporting personalized and data-driven learning strategies.

1. INTRODUCTION

With the emergence of digital learning platforms, educational institutions are now equipped with the ability to capture and analyze large-scale student interaction data. Learning Management Systems (LMS) such as Moodle and Blackboard record students' activities including attendance, quiz participation, time-on-task, and forum interactions, which can be invaluable for

understanding learning behaviors and outcomes [1]. While traditional assessments measure academic achievement at fixed intervals, LMS data provides a continuous stream of behavioral indicators that reflect student engagement in real-time.

Conventional student evaluation methods often overlook patterns in these behavioral indicators. To address this, the field of Educational Data Mining (EDM) has emerged, combining data science techniques with pedagogical insights to predict, understand, and optimize student learning experiences [2], [3]. Machine learning (ML) is a key component of EDM, allowing systems to discover complex, nonlinear relationships between student behavior and academic outcomes [4], [5]. Among ML methods, ensemble approaches like Random Forest are popular for their robustness, accuracy, and interpret-ability [6]. Random Forest Regression, in particular, is a powerful predictive model for continuous variables such as final grades. It works by aggregating predictions from multiple decision trees, each trained on different subsets of the data, thus improving generalization and reducing over-fitting [7]. Additionally, this model offers feature importance metrics that indicate which behavioral inputs—like quiz scores or forum activity—are most influential in predicting academic success [8]. This study applies Random Forest Regression to predict final academic performance based on simulated LMS data, including attendance rate, average quiz scores, LMS usage time, and forum activity. The aim is to answer two key questions: (1) How accurately can academic performance be predicted from behavioral indicators? and (2) Which behaviors contribute most to the prediction model?

The significance of this research lies in its practical implications. By enabling early identification of students who may be under-performing, institutions can design timely interventions and personalize learning support [9], [10]. Furthermore, such predictive analytics can inform instructional design and resource allocation in digital learning environments. The rest of this paper is organized as follows: Section 2 presents the data-set and methods, Section 3 discusses the results, and Section 4 concludes with insights and future work.

2. METHODS

This study uses a dataset of 100 synthetic student records simulating LMS behavioral data. Each record contains five attributes:

(1) Attendance Rate (normalized between 0 and 1),
(2) Average Quiz Score (in percentage),
(3) Weekly Time Spent on LMS (in hours),
(4) Forum Activity (count of posts/replies),
(5) Final Grade (target variable).

The preprocessing phase included checking for missing values, standardizing features using the StandardScaler, and splitting the dataset into training and testing sets using an 80:20 ratio. Feature normalization was essential to prevent scale bias in the model.

A Random Forest Regressor was selected due to its robustness, ability to handle non-linear relationships, and capacity for evaluating feature importance. The model was implemented using the Scikit-learn library [14] with 100 estimators and default parameters. Model performance was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), standard metrics in regression tasks [10].

3. RESULTS AND DISCUSSION

The Random Forest model achieved a MAE of 4.57 and RMSE of 5.90, indicating good predictive accuracy for student final grades. Figure 1 shows the alignment between predicted and actual grades, revealing that most predictions fall close to the diagonal ideal line.

Feature importance analysis (Figure 2) demonstrates that average quiz scores contributed the most to the prediction, followed by attendance rate, LMS usage

time, and forum activity. This aligns with educational theory, where formative assessments and consistent engagement are considered key predictors of academic achievement [3].

The results validate the effectiveness of Random Forest in capturing complex relationships in behavioral data. Such models can be used in academic support systems to monitor student performance in real time and trigger early interventions when necessary.

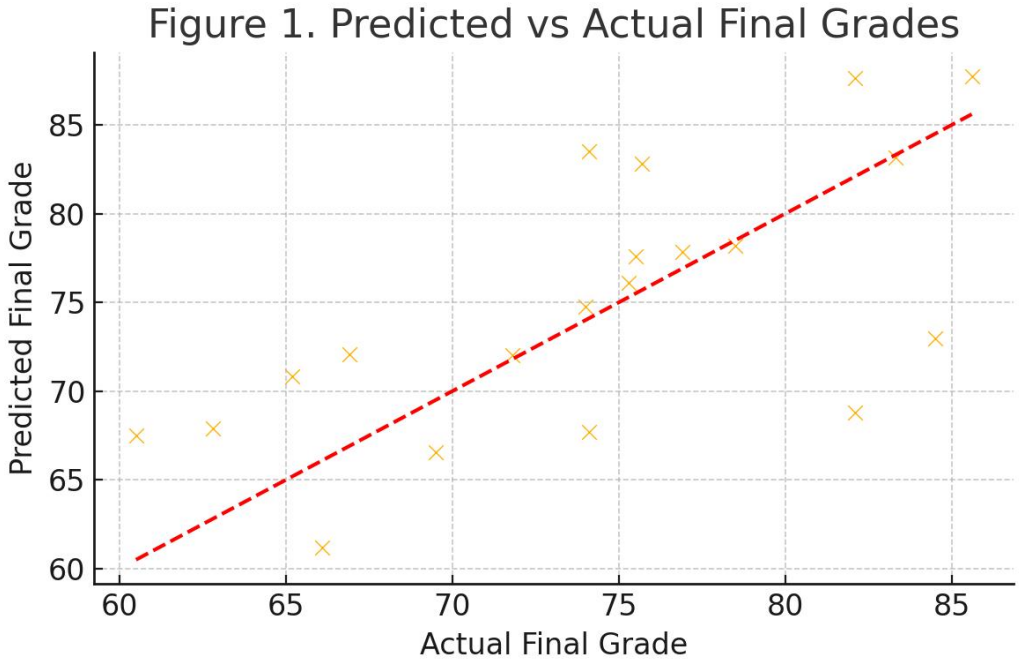

Figure 1. Predicted vs Actual Final Grades

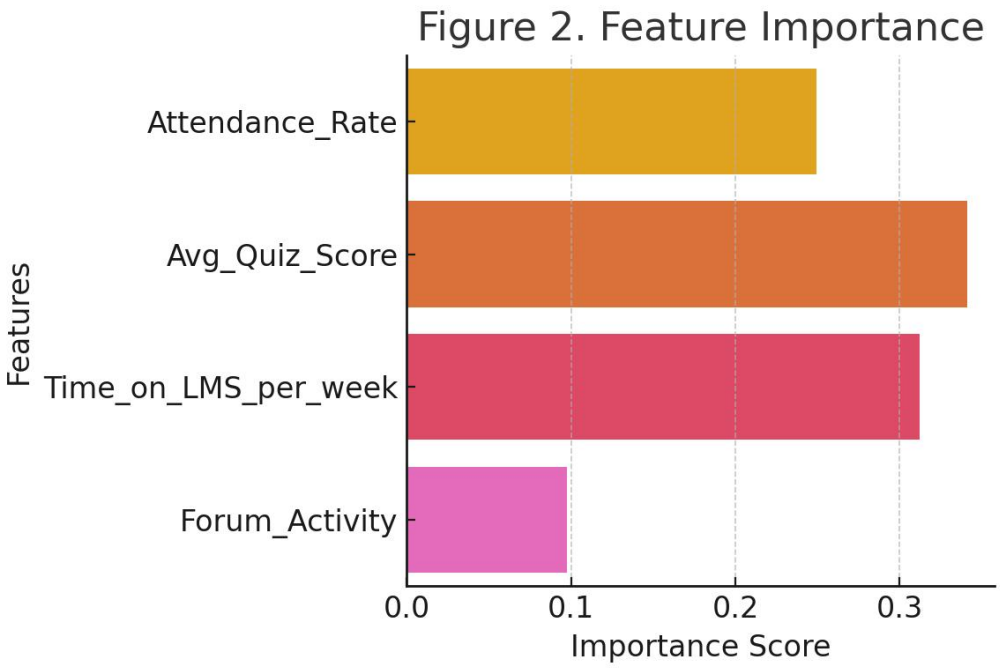

Figure 2. Feature Importance of LMS Features

1. CONCLUSION

This paper demonstrated that student academic performance can be predicted effectively using behavioral data from LMS platforms and Random Forest Regression. The model achieved acceptable accuracy, and its interpretability makes it useful for informing educational interventions. In future work, real-world datasets could be used to validate the findings, and other machine learning models including classification and deep learning approaches could be explored for comparison. Real-time deployment in LMS systems could also be considered for continuous monitoring of student engagement and risk detection.

6. REFERENCES

[1] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert Systems with Applications, vol. 33, no. 1, pp. 135–146, 2007.
[2] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," Applied Artificial Intelligence, vol. 18, no. 5, pp. 411–426, 2004.
[3] R. Baker and P. Inventado, "Educational data mining and learning analytics," in Learning Analytics, Springer, 2014, pp. 61–75.
[4] H. Drachsler and W. Greller, "Privacy and analytics–it's a DELICATE issue," in LAK '16, ACM, 2016.
[5] T. Mitchell, Machine Learning, McGraw-Hill, 1997.
[6] A. Jayaprakash et al., "Early alert of academically at-risk students: An open source analytics initiative," Journal of Learning Analytics, vol. 1, no. 1, pp. 6–47, 2014.
[7] R. Asif et al., "Predicting student academic performance using data mining methods," International Journal of Computer Science and Engineering, vol. 7, no. 5, pp. 245–253, 2015.
[8] Y. Chi et al., "Enabling personalized learning paths with learning analytics," in Proc. IEEE ICALT, 2017.
[9] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[10] J. Brownlee, Regression Metrics for Machine Learning, Machine Learning Mastery, 2020. [Online]. Available: https://machinelearningmastery.com/regression-metrics-for-machine-learning/
[11] S. You, Q. Gu, and Y. Ding, "A survey on interpretability of machine learning," arXiv preprint arXiv:2301.00351, 2023.
[12] G. Siemens and R. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in LAK '12, ACM, 2012.
[13] H. Alyahyan and M. Düştegör, "Predicting academic success in higher education: literature review and best practices," International Journal of Educational Technology in Higher Education, vol. 17, no. 1, 2020.
[14] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.