USING ARTIFICIAL INTELLIGENCE TO PREDICT AND ANALYZE UNIVERSITY STUDENTS' ACADEMIC PERFORMANCE IN DIFFERENT COURSES

Jalolova Mahbubaxon Shavkatjon qizi
Master of Technology
Sharda University Uzbekistan
https://doi.org/10.5281/zenodo.24278468

Abstract. This paper discusses the use of artificial intelligence and machine learning methods to predict university students' academic performance and identify potential academic risk factors at an early stage. The analysis is based on the Open University Learning Analytics Dataset (OULAD), an openly available dataset compiled by the Open University in the United Kingdom. The statistics presented in this paper are drawn from the published dataset and academic literature; model accuracy metrics have not been fabricated, as no separate experiment was conducted.
Keywords: artificial intelligence, machine learning, academic performance, student outcomes, learning analytics, OULAD.

Introduction

The expansion of digital learning environments has led universities to generate large volumes of data about student activity. Logins, views of learning materials, assignment submissions, test results, and submission times recorded on e-learning platforms describe students' participation in the learning process. Systematic analysis of these data can help identify declining academic performance or the risk of not completing a course in advance.

Traditional assessment often records a student's result at the end of a semester. An artificial intelligence-based approach can analyze indicators collected throughout the semester and provide early warnings to instructors and academic advisers. However, a prediction should not be treated as a final judgment about a student; rather, it should be interpreted as a decision-support tool that helps identify the need for additional educational support.

The statistical basis of this paper is the OULAD dataset introduced by Kuzilek, Hlosta, and Zdrahal in the journal Scientific Data in 2017. The dataset links anonymized demographic information about students enrolled in Open University courses with registration records, assessment results, and daily activity logs from the Virtual Learning Environment (VLE) platform [1].

General statistical profile of the Open University Learning Analytics Dataset (OULAD)
Table 1

Indicator	Reported value in OULAD	Note
Student-course records	32 593	One student may enroll in multiple courses
Unique students	28 785	After accounting for repeated course records
Course presentations	22	Across 7 modules
VLE activity	10 655 280	Aggregated daily activity

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

records		
Assessment records	173 912	Assignment and assessment results
Assessment types/items	206	Records in the Assessments table

The figures in Table 1 are reported in the scientific description of the OULAD dataset. A "student-course record" is not the same as a unique individual: one student may be enrolled in more than one module. Therefore, the figure of 32,593 should not be interpreted as the number of unique students [1].

The features used to predict a student's final outcome can be divided into several groups: (1) academic features - continuous assessment scores, assignment results, and previous attempts; (2) engagement features - logins to the VLE platform, views of learning resources, and use of tests and forums; (3) temporal features - assignment submission deadlines and week-by-week changes in activity; and (4) contextual features - course type and workload. Demographic features should be used only where there is a lawful basis and fairness considerations are taken into account.

Comparative overview of artificial intelligence algorithms used to predict student academic performance

Table 2

Algorithm	Function	Advantage	Limitation
Logistic regression	Estimates the outcome as a probability	Easy to interpret; useful as a baseline model	Limited ability to represent complex nonlinear relationships
Decision tree	Classifies based on rules	Easy to interpret and visualize	Prone to overfitting
Random Forest	Combines the results of multiple trees	Robust; can identify complex relationships	More difficult to explain the reason behind an individual decision
Gradient Boosting	Classifies using sequentially boosted models	Often has strong predictive capability	Requires parameter tuning and monitoring

When building a practical model, the data are first cleaned, missing values are checked, categorical features are encoded, and chronological order is preserved. The data are then divided into training, validation, and test sets. Features that become available only after the final outcome must not be included when predicting a student's final result. Otherwise, data leakage occurs and the model's performance may appear artificially high compared with its performance in real-world conditions.

Model performance should not be evaluated using accuracy alone. Recall is important for avoiding missed at-risk students; precision helps control false alerts; F1-score balances these two measures; and ROC-AUC or PR-AUC can be used to assess class discrimination. When classes are imbalanced, PR-AUC and class-specific recall are particularly important.

In the published literature, the four OULAD final outcome categories - Fail, Withdrawn, Pass, and Distinction - have been combined into two broader classes. Fail and Withdrawn are treated as the "at-risk group," while Pass and Distinction are treated as "successful outcomes." Under this classification, 17,208 records belong to the at-risk group and 15,385 records belong to the successful group [2].

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

Distribution of students' final academic outcomes in the OULAD dataset

Table 3

Final group	Number of records	Share
Successful (Pass + Distinction)	15 385	47.2 %
At-risk group (Fail + Withdrawn)	17 208	52.8 %
Total	32 593	100.0 %

The percentages were calculated using the following formula: number of records in a group / total number of records × 100. For example, 15,385 / 32,593 × 100 ≈ 47.2%. Thus, in this two-class classification, the share of the at-risk group is higher than that of the successful group. This result highlights the importance of early-warning mechanisms in the learning process; however, it should not be interpreted as a general rate applicable to all universities or all students.

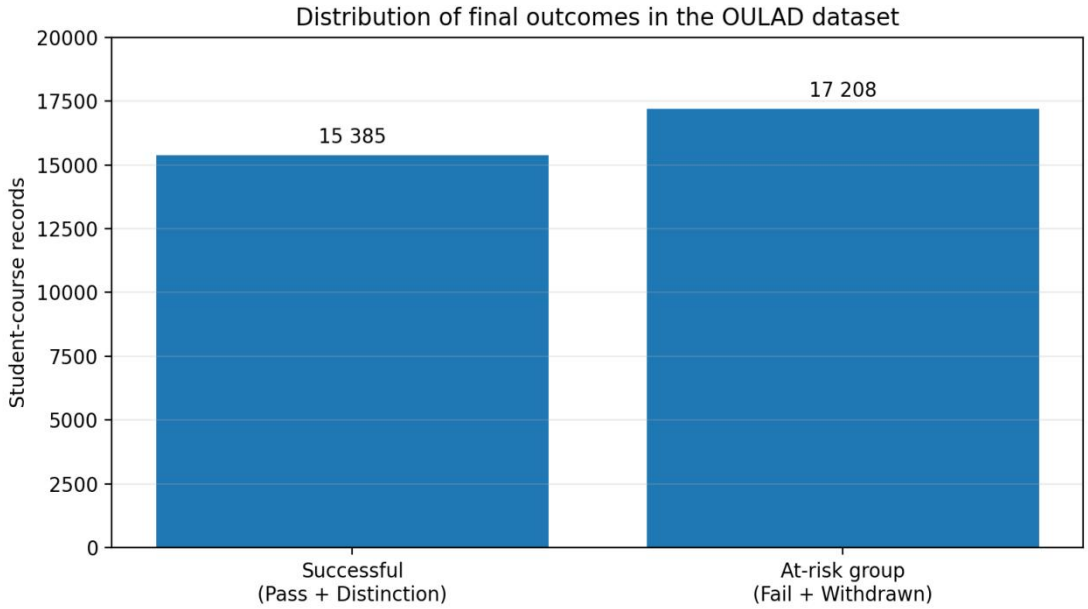


Figure 1. Distribution of OULAD outcomes across the two groups. Source: Chart prepared by the author based on the published figures in [2].

The following step-by-step model is proposed for implementation in a university setting:

Stages of developing and implementing an artificial intelligence-based model for predicting student academic performance

Table 4

Stage	Activity to be performed	Expected outcome
1. Data collection	Obtain authorized data from the LMS, electronic gradebook, and assignment systems	A unified analytical dataset
2. Data preparation	Check duplicate records, missing values, and timestamps	Quality-checked data
3. Feature engineering	Calculate weekly activity, assignment submission delays, and score dynamics	Meaningful predictive features

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

4. Model training	Compare logistic regression and tree-based algorithms	A tested predictive model
5. Evaluation	Evaluate precision, recall, F1, ROC-AUC/PR-AUC, and calibration	Assessment of model reliability
6. Pedagogical intervention	Offer individual support through an instructor or tutor	Timely support

The model's output should not be displayed in a student's personal account as a definitive label such as "unsuccessful." For example, if the system detects declining platform activity, late assignment submissions, and falling continuous assessment scores over several weeks, it may send an alert about potential risk to the responsible tutor. The final decision should be made by an instructor or academic adviser after considering the context.

To protect personal data, it is necessary to minimize the data collected, anonymize or pseudonymize identifiers, restrict access rights, define retention periods, and examine model errors across different student groups. Regular audits should be conducted to ensure that demographic features do not lead to discriminatory outcomes.

Conclusion

In conclusion, artificial intelligence and machine learning algorithms make it possible to analyze university students' academic performance not only through final grades but also through daily engagement and interim results during the learning process. The OULAD database, with 32,593 student-course records and more than 10.65 million VLE activity records, is an important open data source for testing this approach. In the two-class classification reported in the literature, 52.8% of records belong to the at-risk group and 47.2% to the successful-outcome group. These proportions demonstrate the need to account for class imbalance, evaluate models using multiple metrics, and pay particular attention to recall when identifying at-risk students. Logistic regression can be used as an interpretable baseline model, while Random Forest and gradient boosting can be applied to identify more complex relationships. Therefore, AI-based learning analytics can help identify students' academic difficulties at an early stage and plan individual educational support. However, a model's effectiveness can be confirmed only after it has been evaluated using real university data, a correctly time-separated test set, and appropriate fairness and privacy safeguards.

References:

1. Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset // Scientific Data. – 2017. – Vol. 4. – Article 170171. – DOI: 10.1038/sdata.2017.171. – URL: https://doi.org/10.1038/sdata.2017.171
2. Students' Academic Performance and Engagement Prediction in a Virtual Learning Environment Using Random Forest with Data Balancing // Sustainability. – 2022. – Vol. 14, No. 22. – Article 14795. – URL: https://www.mdpi.com/2071-1050/14/22/14795
3. Open University Learning Analytics Dataset (OULAD). The Open University, Knowledge Media Institute. – URL: https://analyse.kmi.open.ac.uk/open_dataset
4. Learning Analytics: Analysis of Methods for Online Assessment // Applied Sciences. – 2022. – Vol. 12, No. 18. – Article 9296. – URL: https://www.mdpi.com/2076-3417/12/18/9296

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.