PREDICTING ACADEMIC SUCCESS WITH MACHINE LEARNING: A SIMULATION-BASED ANALYTICS FRAMEWORK FOR A REGIONAL UNIVERSITY IN UZBEKISTAN

Kodirov Abdujabbor Sirojiddin Ugli
email: abdujabborqodirov@gmail.com
ORCID iD: 0009-0004-3929-5480
International Joint Degree Program, Termez State University and Faculty of Mathematics and Physics Education, Universitas Pendidikan Indonesia

Abstract: *Higher education institutions in developing regions increasingly generate large volumes of student data, yet most continue to rely on manual, retrospective evaluation rather than predictive analytics. This study examines whether supervised machine learning can support early identification of academically at-risk students at Termez State University, Uzbekistan, a context in which empirical evidence on educational analytics remains scarce. A simulated institutional dataset of 2,450 undergraduate records, constructed to reflect realistic academic, attendance, demographic, and digital-engagement parameters, was used to train and compare four classifiers: Logistic Regression, Support Vector Machine, Random Forest, and Artificial Neural Network. Academic success was operationalised as a binary outcome (GPA ≥ 2.50 versus GPA < 2.50). Descriptive and correlation analyses preceded model training on an 80/20 split, with performance evaluated via accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices. Random Forest achieved the strongest performance (accuracy = 0.89, F1 = 0.86, ROC-AUC = 0.91), outperforming the Artificial Neural Network (0.87), Support Vector Machine (0.85), and Logistic Regression (0.82). Final examination score, attendance rate, assignment score, and previous GPA were the most influential predictors, while demographic variables contributed minimally. Approximately 24% of students were classified as academically at risk. The findings suggest that ensemble learning methods offer a feasible, interpretable foundation for early-warning systems in resource-constrained regional universities, while underscoring the need for real institutional data, broader psychosocial variables, and explicit governance safeguards before operational deployment.*

Keywords: *learning analytics; machine learning; academic performance prediction; Random Forest; early warning systems; higher education; Uzbekistan; educational data mining*

1. INTRODUCTION

The digitalisation of higher education has generated unprecedented volumes of student data, spanning attendance logs, assessment records, learning-management-system (LMS) interactions, and demographic profiles. Machine learning (ML) methods increasingly convert these records into predictive intelligence that supports student retention, early intervention, and institutional planning (Romero & Ventura, 2020; Baker & Hawn, 2021). International evidence shows that supervised algorithms such as Random Forest, Support Vector Machines, and neural networks can classify at-risk students with strong accuracy well before final outcomes are known

(Costa et al., 2017; Al-Shabandar et al., 2019), enabling proactive counselling rather than reactive remediation.

Uzbekistan's higher education sector is undergoing extensive digital transformation under the "Digital Uzbekistan 2030" programme, with universities adopting electronic academic-management systems and online learning platforms. However, this digitalisation has so far centred on administrative record-keeping rather than analytical or predictive use of institutional data (Abidov et al., 2023). Termez State University, a leading public university in southern Uzbekistan, exemplifies this pattern: it accumulates extensive student records, yet decisions about academic risk, curriculum quality, and retention continue to rely on manual documentation and subjective judgement rather than systematic modelling. Without predictive capability, at-risk students are typically identified only after final examinations, teaching effectiveness is difficult to benchmark, and resource allocation across faculties remains reactive.

Regional scholarship on educational analytics in Central Asia remains limited, with most existing work addressing ICT infrastructure and digital-literacy policy rather than applied predictive modelling (Tursunov, 2021; Sharipov & Tashpulatova, 2022). Consequently, there is little empirical evidence on which student-level indicators best predict academic outcomes in this context, or on which ML algorithms are best suited to the data conditions typical of regional Uzbek universities. This study addresses that gap by developing and comparing four supervised classifiers on a dataset constructed to emulate realistic institutional conditions at Termez State University, with the aim of establishing a feasibility baseline for future analytics-driven student support.

The specific objectives were to (i) identify which academic, behavioural, and demographic variables most strongly predict student academic success; (ii) compare the classification performance of Logistic Regression, Support Vector Machine, Random Forest, and Artificial Neural Network models; and (iii) evaluate the practical implications of the results for early-warning systems and institutional decision-making in a resource-constrained regional university.

2. Literature Review

2.1 Analytics in Higher Education

Learning analytics has evolved from descriptive institutional reporting toward predictive and prescriptive modelling, driven largely by the integration of machine learning into student-information systems (Romero & Ventura, 2020). Universities use analytics to address dropout prediction, performance forecasting, teaching evaluation, and resource planning (Daniel, 2019; Baker & Hawn, 2021). A substantial strand of this literature concerns student retention: educational data mining techniques have repeatedly demonstrated that dropout and academic-risk indicators can be detected well before final results, enabling timely academic counselling and tutoring (Costa et al., 2017; Al-Shabandar et al., 2019). Digitised learning environments, particularly LMS platforms, have expanded the granularity of available behavioural data — login frequency, submission timing, and quiz engagement — which several studies link to both academic outcomes and learning motivation (Papamitsiou & Economides, 2014; Kizilcec & Brooks, 2017).

Despite these benefits, the literature also documents persistent implementation challenges: data-privacy and consent concerns, uneven institutional capacity, and the risk that predictive

labels may inadvertently harm student motivation if not carefully governed (Prinsloo & Slade, 2017; Baker & Hawn, 2021). These concerns are especially salient for universities in developing regions, where computational infrastructure and trained analytics personnel remain limited (Ifenthaler & Yau, 2020).

2.2 Machine Learning Approaches to Student Performance Prediction

Comparative studies consistently show that ensemble methods, particularly Random Forest, perform competitively against both simpler baseline models and more complex neural architectures for student outcome classification, while retaining an interpretability advantage through feature-importance ranking (Hussain et al., 2020; Lu et al., 2021). Logistic Regression remains a common interpretable baseline (Costa et al., 2017), while Support Vector Machines are favoured for their robustness in higher-dimensional feature spaces (Al-Shabandar et al., 2019). More recent work extends this comparison to gradient-boosting methods: Pek et al. (2023) reported that ensemble tree-based classifiers substantially reduced missed-risk cases relative to simpler models, and Mduma (2023) showed that data-balancing techniques materially improve dropout-prediction recall when outcome classes are skewed, a condition that also characterises the present dataset (24% at-risk versus 76% successful). Explainable-AI techniques are increasingly paired with these models: Nnadi et al. (2024) applied post-hoc explanation methods alongside Random Forest to interpret adaptability predictions, reinforcing the argument that predictive accuracy alone is insufficient for institutional adoption without transparent reasoning that academic staff can act upon.

2.3 Educational Analytics in the Uzbek Context and Research Gap

Uzbek higher education research has predominantly examined digitalisation policy, ICT adoption, and curriculum modernisation rather than applied predictive analytics (Tursunov, 2021; Sharipov & Tashpulatova, 2022). Empirical studies using machine learning on Uzbek or wider Central Asian student data remain scarce, and no published work has compared multiple ML algorithms for academic-risk prediction within this specific institutional setting. In addition, existing regional literature relies heavily on descriptive statistics or small-sample surveys rather than validated predictive models with standard performance metrics such as F1-score or ROC-AUC (Ibragimov, 2020). Ethical and fairness considerations, increasingly central to the international learning-analytics literature (Memarian & Doleck, 2023), are similarly underexplored in the regional context. This study addresses these combined gaps by applying, comparing, and interpreting four supervised ML models within a Termez State University-oriented dataset, explicitly incorporating fairness and interpretability considerations alongside predictive performance.

3. Materials and Methods

3.1 Research Design

This study adopted a quantitative, supervised machine learning design to evaluate the predictive relationship between student-level variables and academic outcomes. The design comprised three stages: (i) descriptive and correlational analysis to characterise the dataset and identify candidate predictors; (ii) training and comparison of four supervised classifiers; and (iii) interpretation of model outputs — including feature importance and confusion-matrix analysis — for their institutional implications.

3.2 Dataset

Because direct extraction of live institutional records from Termez State University's information systems was not feasible within the scope of this study, the dataset was constructed to emulate realistic academic, attendance, demographic, and digital-engagement conditions reported in institutional documentation and comparable regional universities. This approach follows established practice in early-stage feasibility studies, where a parameterised, literature-informed synthetic dataset is used to establish whether a proposed analytics pipeline and model selection are viable before an operational deployment using verified administrative data is pursued. Consequently, the results reported here should be interpreted as a model-based feasibility analysis rather than an evaluation of a live institutional system; this constraint is discussed further in the Limitations section.

The dataset comprised 2,450 undergraduate student records, retained after cleaning from an initial 2,680, spanning four academic years and four faculty groupings: Economics and Business (n = 650), Education and Pedagogy (n = 550), Social Sciences and Humanities (n = 500), and STEM programmes (n = 400), plus other programmes (n = 350, summing with rounding across the cleaned sample). Variables included academic performance indicators (GPA, final examination score, assignment score, quiz score), attendance metrics (attendance rate, absence count), demographic attributes (gender, age, year level, faculty, admission type), and digital-engagement indicators (LMS login frequency, submission punctuality, online quiz attempts, available for 2,120 of 2,450 records). The outcome variable, academic success status, was binary: students with a semester GPA ≥ 2.50 were coded as academically successful (n = 1,862, 76.0%), and those below this threshold as academically at risk (n = 588, 24.0%).

3.3 Data Preprocessing

Preprocessing addressed duplication, missingness, and scale heterogeneity. Duplicate records (n = 74) identified via student and semester registration codes were removed, as were records missing GPA (n = 96) or attendance data (n = 60), since these variables were essential to the outcome definition. Missing LMS-activity values (n = 330) were imputed using the median, given uneven platform adoption across faculties. Categorical variables (gender, faculty, admission type, year level) were label-encoded, and continuous variables were standardised, a step of particular importance for the Support Vector Machine and Artificial Neural Network models, which are sensitive to feature-scale differences.

3.4 Machine Learning Models

Four supervised classifiers were trained and compared: (i) Logistic Regression, selected as an interpretable linear baseline; (ii) Support Vector Machine (SVM), chosen for robustness in multidimensional, non-linearly separable feature spaces; (iii) Random Forest, an ensemble of decision trees offering variance reduction and native feature-importance estimation; and (iv) Artificial Neural Network (ANN), included to assess whether non-linear deep-learning representations improved on tree-based and linear approaches. All models used the same predictor set — final exam score, assignment score, quiz score, attendance rate, absence count, previous GPA, faculty, year level, admission type, and LMS login frequency — to ensure a fair comparison. The dataset was split 80/20 into training (n = 1,960) and testing (n = 490) subsets.

3.5 Evaluation Metrics

Model performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC, complemented by confusion-matrix analysis and Random Forest feature-importance rankings. Given the class imbalance (76% successful versus 24% at-risk), recall for the at-risk class was treated as a particularly important indicator, since false negatives — at-risk students misclassified as successful — carry the greatest practical cost in an early-warning context.

3.6 Ethical Considerations

Although the dataset used in this study was simulated rather than extracted from live records, the research was designed as though operating under the ethical constraints that would apply to real institutional data: identifiers would be replaced with coded values, access restricted to research purposes, and institutional permission obtained before any extraction of genuine student records. Model performance was also considered across demographic subgroups (gender, admission type) to monitor for potential algorithmic bias, consistent with recommended practice in the fairness-in-learning-analytics literature (Baker & Hawn, 2021; Memarian & Doleck, 2023).

4. Results

4.1 Descriptive Characteristics

Across the cleaned sample (n = 2,450), the mean GPA was 2.87 (SD = 0.54, range 1.20–4.00), the mean attendance rate was 82.4% (SD = 11.8), and the mean final examination score was 71.8 out of 100 (SD = 12.6). Mean assignment and quiz scores were 74.5 (SD = 13.2) and 69.7 (SD = 14.1) respectively, and mean LMS login frequency was 18.6 sessions (SD = 10.4). Faculty-level comparison showed that STEM programmes had the lowest average GPA (2.74) and the highest proportion of at-risk students (126 of 400, 31.5%), compared with Education and Pedagogy, which had the highest average GPA (2.95) and the lowest at-risk proportion (108 of 550, 19.6%).

4.2 Correlation with Academic Outcome

Final examination score showed the strongest association with GPA (r = 0.78), followed by assignment score (r = 0.71). Attendance rate was moderately positively correlated with GPA (r = 0.64), while absence count was moderately negatively correlated (r = −0.61). Quiz score (r = 0.59) and LMS login frequency (r = 0.46) showed weaker but still positive associations, and age showed negligible correlation (r = 0.08).

Variable	Correlation with GPA	Strength
Final exam score	0.78	Strong positive
Assignment score	0.71	Strong positive
Attendance rate	0.64	Moderate positive
Quiz score	0.59	Moderate positive
LMS login frequency	0.46	Moderate positive
Absence count	−0.61	Moderate negative
Age	0.08	Weak positive

Table 1. Correlation between predictor variables and GPA.

4.3 Model Performance Comparison

Table 2 reports classification performance for the four models on the held-out test set (n = 490). Random Forest achieved the highest scores across all metrics (accuracy = 0.89, precision = 0.87, recall = 0.85, F1 = 0.86, ROC-AUC = 0.91), followed by the Artificial Neural Network (accuracy = 0.87, ROC-AUC = 0.89), Support Vector Machine (accuracy = 0.85), and Logistic Regression (accuracy = 0.82).

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.82	0.79	0.76	0.77	0.84
Support Vector Machine	0.85	0.82	0.80	0.81	0.87
Random Forest	0.89	0.87	0.85	0.86	0.91
Artificial Neural Network	0.87	0.85	0.82	0.83	0.89

Table 2. Machine learning model performance comparison (test set, n = 490).

4.4 Confusion Matrix and Feature Importance

The Random Forest confusion matrix (Table 3) shows that 361 of 379 actually successful students were correctly classified, alongside 75 of 111 actually at-risk students. Eighteen successful students were misclassified as at-risk (false positives), while 36 at-risk students were misclassified as successful (false negatives) — the more consequential error type for an early-warning application.

Actual / Predicted	Predicted successful	Predicted at-risk
Actual successful	361	18
Actual at-risk	36	75

Table 3. Confusion matrix for the Random Forest model.

Feature-importance analysis from the Random Forest model (Table 4) identified final exam score (0.24), attendance rate (0.19), assignment score (0.17), previous GPA (0.15), and LMS login frequency (0.10) as the five strongest predictors. Demographic variables — year level (0.03), faculty (0.02), admission type (0.01), and gender (0.01) — contributed minimally to model predictions.

Rank	Variable	Importance score
1	Final exam score	0.24
2	Attendance rate	0.19
3	Assignment score	0.17
4	Previous GPA	0.15
5	LMS login frequency	0.10
6	Quiz score	0.08
7	Year level	0.03
8	Faculty	0.02
9	Admission type	0.01

Rank	Variable	Importance score
10	Gender	0.01

Table 4. Random Forest feature-importance ranking.

5. Discussion

The results indicate that academic success in this dataset was driven primarily by continuous academic engagement — examination performance, assignment completion, attendance, and prior GPA — rather than by demographic background. This pattern is consistent with international evidence that behavioural and assessment-based indicators outperform demographic variables in predicting student outcomes (Costa et al., 2017; Hussain et al., 2020), and it carries a favourable equity implication: predictive support strategies at Termez State University could reasonably be framed around study behaviour rather than fixed student characteristics, reducing the risk that predictive labelling reinforces demographic disadvantage.

Random Forest's superior performance relative to Logistic Regression, SVM, and ANN aligns with a broader pattern in the educational-ML literature, where ensemble tree-based methods balance accuracy with interpretability more effectively than either simple linear baselines or opaque neural architectures (Hussain et al., 2020; Pek et al., 2023). The ANN's marginally lower accuracy despite its architectural complexity echoes a recurring finding that additional model capacity does not automatically translate into better classification when the underlying feature set is already strongly predictive and the dataset size is moderate. For an institution such as Termez State University, where analytics capacity and technical staffing are still developing, Random Forest's relative interpretability — via directly inspectable feature importance rather than post-hoc explanation methods — is a practical as well as statistical advantage (Nnadi et al., 2024).

The confusion-matrix results highlight an operationally important asymmetry: of the two error types, false negatives (at-risk students predicted successful) were more frequent than false positives, and represent the more harmful outcome, since these are precisely the students who would fail to receive timely support. This mirrors observations elsewhere that raw accuracy can understate a model's practical adequacy for early-warning use cases, where recall on the minority (at-risk) class deserves particular weight (Mduma, 2023). Future refinement of this framework should therefore prioritise class-imbalance-aware techniques — resampling, class weighting, or threshold adjustment — over further accuracy optimisation alone.

The faculty-level disparity, with STEM programmes showing the lowest average GPA and the highest at-risk proportion, suggests that a single institution-wide risk threshold may not be equally well calibrated across disciplines. This is consistent with international findings that risk indicators are context-dependent: attendance may matter more in laboratory-based instruction, while assignment punctuality may be more diagnostic in humanities-oriented programmes. A single global model may therefore benefit from faculty-specific calibration or covariates in future iterations.

More broadly, these findings extend the empirical base for educational analytics beyond the predominantly Western and East Asian contexts that dominate the literature (Al-Shabandar et al., 2019; Lu et al., 2021), and provide an initial, regionally grounded reference point for Central Asian higher education. However, because the underlying dataset was constructed rather

than extracted from live institutional systems, these results should be read as a demonstration of technical and analytical feasibility rather than as validated operational evidence — a distinction that is central to interpreting the study's contribution appropriately.

6. Conclusion

This study developed and compared four supervised machine learning models to predict academic success on a simulated dataset constructed to reflect institutional conditions at Termez State University. Random Forest provided the strongest and most interpretable classification performance (accuracy = 0.89, F1 = 0.86, ROC-AUC = 0.91), with final examination score, attendance rate, assignment score, and previous GPA emerging as the dominant predictors, and demographic variables contributing minimally. Approximately one in four students in the dataset fell below the academic-success threshold, underscoring the practical relevance of early-warning capability for regional universities that currently rely on retrospective, manual evaluation. These findings support the feasibility of ensemble-based predictive analytics as a foundation for future institutional early-warning systems in Uzbekistan's higher education sector, while making clear that operational deployment would require verified institutional data, explicit data-governance policy, and continued human oversight of any resulting academic decisions.

Limitations

Several limitations qualify these findings. First, and most importantly, the dataset was constructed to emulate realistic institutional parameters rather than extracted from Termez State University's live administrative systems; results should therefore be interpreted as a model-based feasibility analysis, and replication with verified institutional records is a necessary next step before any operational claims can be made. Second, several variables plausibly relevant to academic outcomes — socio-economic status, psychological wellbeing, family support, and internet-access quality — were not available and could not be modelled. Third, LMS-derived indicators may under-represent true digital engagement where platform adoption is uneven across faculties. Fourth, the study is confined to a single institution, limiting generalisability to other Uzbek or Central Asian universities without further validation. Fifth, only four algorithms were compared; more recent gradient-boosting and explainable-AI approaches (e.g., XGBoost, LightGBM, SHAP-based interpretation) were outside the present scope. Sixth, feature-importance results describe predictive association rather than causal mechanism; for example, attendance may proxy for motivation or preparation rather than independently causing higher GPA. Finally, no operational deployment, governance framework, or bias-audit protocol was implemented, and these remain prerequisites for responsible institutional use.

REFERENCES:

Al-Shabandar, R., Hussain, A., Keight, R., Laws, A., & Radi, N. (2019). Machine learning approaches to predict student academic performance. Applied Computing and Informatics, 15(1), 1–12. https://doi.org/10.1016/j.aci.2018.08.002

Arnold, K., & Pistilli, M. (2012). Course signals at Purdue: Using learning analytics to increase student success. LAK '12 Proceedings, 267–270. https://doi.org/10.1145/2330601.2330666

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. International Journal of Artificial Intelligence in Education, 32(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Baker, R., & Hawn, A. (2021). Algorithmic fairness and machine learning in education. Journal of Learning Analytics, 8(3), 6–24. https://doi.org/10.18608/jla.2021.7437

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior, 73, 247–256. https://doi.org/10.1016/j.chb.2017.01.047

Costa, E. B., Fonseca, B., Santana, M., de Araújo, F., & Rego, J. (2017). Evaluating student dropout risk using machine learning techniques. Journal of Educational Computing Research, 57(2), 364–392. https://doi.org/10.1177/0735633118757013

Creswell, J. W., & Creswell, J. D. (2023). Research design: Qualitative, quantitative, and mixed methods approaches (6th ed.). SAGE Publications.

Daniel, B. (2019). Big data and analytics in higher education: Opportunities and challenges. British Journal of Educational Technology, 50(1), 239–250. https://doi.org/10.1111/bjet.12723

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2020). Student academic performance prediction using supervised learning techniques. IEEE Access, 8, 136–152. https://doi.org/10.1109/ACCESS.2020.2965271

Ifenthaler, D., & Yau, J. (2020). Utilising artificial intelligence to support teacher data analytics literacy. Computers in Human Behavior, 113, 106–122. https://doi.org/10.1016/j.chb.2020.106506

Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. Journal of Learning Analytics, 1(1), 6–47. https://doi.org/10.18608/jla.2014.11.3

Kizilcec, R. F., & Brooks, C. (2017). Personalized support improves execution and course completion in MOOCs. Journal of Learning Analytics, 4(1), 24–41. https://doi.org/10.18608/jla.2017.41.3

Lu, H., Li, Y., & Chen, M. (2021). Deep learning for education analytics and prediction. IEEE Access, 9, 40334–40351. https://doi.org/10.1109/ACCESS.2021.3064320

Mduma, N. (2023). Data balancing techniques for predicting student dropout using machine learning. Data, 8(3), 49. https://doi.org/10.3390/data8030049

Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. Computers and Education: Artificial Intelligence, 5, 100152.

Nnadi, L. C., Watanobe, Y., Rahman, M. M., & John-Otumu, A. M. (2024). Prediction of students' adaptability using explainable AI in educational machine learning models. Applied Sciences, 14(12), 5141. https://doi.org/10.3390/app14125141

Papamitsiou, Z., & Economides, A. (2014). Learning analytics and educational data mining in practice. Educational Technology & Society, 17(4), 49–64.

Pek, R. Z., Ozyer, S. T., Elhage, T., Ozyer, T., & Alhajj, R. (2023). The role of machine learning in identifying students at-risk and minimizing failure. IEEE Access, 11, 1224–1243. https://doi.org/10.1109/ACCESS.2022.3232984

Prinsloo, P., & Slade, S. (2017). An ethics framework for learning analytics. British Journal of Educational Technology, 48(5), 153–170. https://doi.org/10.1111/bjet.12261

Romero, C., & Ventura, S. (2020). Educational data mining and machine learning: A decade review. WIREs Data Mining and Knowledge Discovery, 10(3), e1355. https://doi.org/10.1002/widm.1355

Sharipov, J., & Tashpulatova, D. (2022). Education digitalisation in Uzbekistan: Current issues and future opportunities. Central Asian Journal of Education and Technology, 3(2), 55–68.

Note. References by Memarian and Doleck (2023), Mduma (2023), Nnadi et al. (2024), and Pek et al. (2023) are suggested additions (2022–2026) proposed by the writer to strengthen currency on explainability, class-imbalance handling, and fairness in educational machine learning; they were not cited in the original thesis. All other references were drawn from the original thesis reference list.