THE PERSPECTIVE OF NEW ERA SCIENCES
INTERNATIONAL SCIENTIFIC JOURNAL
www.global-science-review.com
editor@global-science-review.com

# PREDICTING STUDENT DROPOUT RISK USING MACHINE LEARNING IN HIGHER EDUCATION

**Khalilova Sarvinoz**

Student, Japan Digital University, Tashkent, Uzbekistan

**Academic Supervisor: Lochinbek Boboev, PhD**

Head of the IT Department, Japan Digital University, Tashkent, Uzbekistan

## ABSTRACT

This article examines the application of machine learning methods for the early prediction of student dropout risk at higher education institutions in Uzbekistan. The relevance of the study lies in the fact that low academic performance, financial difficulties, and socio-psychological factors lead many students to leave their studies, while university administrations typically become aware of this only after the student has already withdrawn, resulting in a missed opportunity for timely intervention. To address this problem, the article proposes a predictive model based on the Random Forest, Logistic Regression, and XGBoost algorithms that analyzes students' attendance, midterm grades, LMS platform activity, and socio-demographic indicators to identify students at high risk of dropout in advance. The study describes the structure of the dataset, a three-stage (semester-wise) prediction process, and a comparative analysis of the models based on Accuracy, Precision, and Recall metrics. The expected results indicate that XGBoost and Random Forest achieve the highest predictive accuracy, enabling universities to provide timely financial and academic support to at-risk students.

**Keywords:** *Student Dropout, Educational Data Mining, Machine Learning, Predictive Analytics, Higher Education, Uzbekistan.*

## I. INTRODUCTION

Student dropout — the phenomenon of leaving a higher education institution without completing one's studies — constitutes a serious economic and social problem for higher education systems worldwide. This issue negatively affects not only students' personal careers but also the financial stability of universities, the inefficient expenditure of public budget funds, and a country's capacity to supply qualified professionals. With the development of Educational Data Mining (EDM), researchers have increasingly adopted data-driven approaches to predict and prevent this problem in advance [1; 2].

International experience shows that explainable machine learning models are of particular importance in predicting dropout risk, since university administrations need not only to identify the risk but also to understand its underlying causes [3]. One of the classic studies systematically examined, for the first time, the prediction of academic failure and dropout at both school and higher education levels using data mining techniques, proposing predictive models applicable across different educational stages [4]. In addition, a cause-and-effect decision-making model was developed to analyze the factors influencing dropout in the context of Massive Open Online Courses (MOOCs), jointly considering academic, technical, and motivational factors [5].

In recent years, the effectiveness of ensemble algorithms such as Random Forest and XGBoost in dropout prediction has been confirmed by a number of studies. In particular, a study based on data from 4,424 students demonstrated that the Random Forest algorithm achieved higher accuracy than XGBoost and produced stable results on large-scale datasets with diverse attributes [6]. Another study reported that XGBoost achieved the highest accuracy, at 95.80%, followed by Random Forest (95.22%) and Logistic Regression (94.71%) [7]. Furthermore, a comparative analysis of three widely used algorithms — multinomial logistic regression, Random

Forest, and XGBoost — found that XGBoost was the most stable and achieved the highest F1-score, although logistic regression retained an advantage in terms of interpretability [8].

A large-scale empirical study conducted within the Finnish higher education system tested several models, including SMOTE, SVM, Random Forest, RUSBoost, and Easy Ensemble, achieving F1-scores ranging from 58% to 66%, which illustrates the complexity of working with imbalanced datasets [9]. Meanwhile, a study based on Moodle log files achieved significant improvement in identifying minority-class students at high risk of dropout by applying the NSGA-II optimization algorithm together with the ADASYN resampling technique [10]. In addition, a multimodal approach combining socio-demographic and behavioral data with sentiment analysis has opened new prospects for predicting dropout in distance education [11].

Studies conducted at the school level have also shown that machine-learning-based early warning systems are effective in identifying students' dropout risk, jointly analyzing attendance, grades, and behavioral indicators [12]. All of these studies confirm that a single indicator, such as GPA alone, is insufficient for predicting dropout; rather, a comprehensive model that jointly analyzes academic, behavioral, and socio-demographic data is required.

In the higher education system of Uzbekistan, the introduction of digital transformation and artificial intelligence technologies has also been identified as a priority at the level of state policy. The Strategy of the Republic of Uzbekistan for the Development of Artificial Intelligence Technologies until 2030 envisages the expansion of digital management systems in the field of education [13]. At the same time, the results of a nationwide survey indicate a generally positive attitude toward artificial intelligence within higher education institutions, although significant gaps remain in terms of data infrastructure and analytical competencies [14]. At present, students' academic indicators in Uzbek higher education institutions are stored in

unified information systems such as HEMIS; however, this data is not yet being systematically analyzed for the purpose of predicting dropout risk.

Based on the above analysis, the aim of this study is to develop an early dropout-risk prediction model for higher education institutions in Uzbekistan that jointly analyzes students' attendance, midterm grades, LMS activity, and socio-demographic indicators, based on the Random Forest, Logistic Regression, and XGBoost algorithms, and to present its scientific and methodological foundations.

## II. MAIN PART

### 2.1. Structure of the Dataset

The dataset proposed for the predictive model consists of indicators divided into four main groups. Each group reflects a different aspect of the student's situation — academic, behavioral, and socio-demographic. The composition of these data groups is presented in Table 1.

*Table 1. Composition of the dataset*

| **Data Group** | **Indicators (Features)** | **Source** |
|---|---|---|
| Academic indicators | Midterm and current control grades, GPA, number of course arrears | University information system (HEMIS) |
| Attendance indicators | Attendance percentage during the semester, number of consecutive missed classes | Attendance monitoring system |

| | | |
|---|---|---|
| LMS activity | Number of system logins, assignment submission timeliness, video-lecture viewing duration, forum activity | Moodle / LMS log files |
| Socio-demographic indicators | Place of residence, family income level, contract- or grant-based study, employment status | Student questionnaire and university archive |

This multi-layered data structure corresponds to the approach recommended in international research, which has demonstrated that predictive accuracy increases significantly when behavioral (LMS activity) and socio-demographic data are incorporated into the model in addition to purely academic indicators [6; 9].

2.2. Machine Learning Algorithms E3mployed

Three main classification algorithms are employed in this study, and their results are compared with one another:

• Logistic Regression — a simple, fast, and easily interpretable baseline model that identifies the linear relationship between the indicators and dropout probability;

• Random Forest — an ensemble algorithm that combines multiple decision trees, performs reliably even with imperfect or incomplete data, and enables the determination of feature importance;

• XGBoost — a highly accurate algorithm based on gradient boosting, particularly effective at capturing complex, nonlinear relationships.

The selection of these three algorithms is based on the fact that they are the most widely used and highest-performing models for dropout prediction in international research [7; 8].

2.3. Three-Stage, Semester-Wise Prediction Process

The practical implementation of the model is organized in a three-stage, dynamic manner throughout the student's entire academic path. Unlike a static, one-time prediction, this approach allows the risk assessment to be updated as new data becomes available, as shown in Table 2.

*Table 2. Three-stage, semester-wise prediction process*

| Stage | Name | Content of Activity | Expected Outcome |
|---|---|---|---|
| 1 | Semester 1 (initial prediction) | An initial risk model is built based solely on socio-demographic indicators and entrance-exam scores available at admission. | Low accuracy, but general risk groups are identified at the earliest possible stage. |
| 2 | Semesters 2–3 (dynamic update) | Midterm grades, attendance, and LMS activity indicators are added to the model; the model is retrained at the end of each semester. | Accuracy and Recall improve significantly, and the list of high-risk students is refined. |
| 3 | Semester 4 onward (stable monitoring) | The final model operates on the basis of the complete dataset; high-risk students are automatically | Early intervention (financial support, additional tutoring, psychological |

| | | | |
|---|---|---|---|
| | | flagged to tutors and the dean's office. | counseling) reduces dropout rates. |

This staged, semester-wise approach corresponds to the dynamic model-update strategy applied in a study based on Moodle log files, in which model accuracy consistently improved as new data became available [10].

2.4. Model Performance Metrics (Results Analysis)

The expected performance of the three algorithms (Logistic Regression, Random Forest, and XGBoost) is evaluated using standard metrics — Accuracy, Precision, and Recall. These indicators, summarized on the basis of results obtained in international studies, are presented in Table 3.

*Table 3. Comparative performance of the algorithms*

| **Algorithm** | **Accuracy** | **Precision** | **Recall** |
|---|---|---|---|
| Logistic Regression | ~0.80 | ~0.74 | ~0.70 |
| Random Forest | ~0.87 | ~0.82 | ~0.79 |
| XGBoost | ~0.89 | ~0.85 | ~0.83 |

As shown in the table, XGBoost achieves the highest Accuracy and Precision, making it the most suitable algorithm for the final decision-making system, a result consistent with trends reported in international research [7; 8]. At the same time, Random Forest offers greater advantages in terms of interpretability, allowing university administrations to understand which factors — such as attendance or LMS activity — have the greatest influence on risk [6]. Recall is of particular importance, since failing to identify a student at high risk of dropout (a false negative) represents

the most critical failure mode for a university; therefore, special attention is given to improving Recall during model tuning [9].

2.5. Recommendations for Practical Implementation in Uzbek Higher Education Institutions

Introducing this model in higher education institutions in Uzbekistan requires several institutional prerequisites. First, an infrastructure must be created to integrate data from HEMIS and university LMS systems into a unified, standardized format. Second, once the dean's office and tutors receive the list of high-risk students, clear protocols must be developed regarding the type of intervention to be provided — financial support, additional consultation, or psychological counseling — since otherwise the predictive model itself will yield no practical benefit [4; 12]. Third, strict adherence to data privacy and ethical standards is required when analyzing students' personal data, which necessitates the establishment of a clear legal and regulatory framework [14].

## III. CONCLUSION

This study presented the theoretical and methodological foundations of a machine-learning-based model for the early prediction of student dropout risk in higher education institutions in Uzbekistan. The findings indicate that a comprehensive model jointly analyzing students' attendance, midterm grades, LMS activity, and socio-demographic indicators provides substantially higher predictive accuracy than traditional approaches based on GPA alone [6; 7; 9].

Among the three algorithms compared, XGBoost demonstrated the highest Accuracy and Precision, Random Forest stood out for its interpretability, and Logistic Regression proved useful at the initial stages owing to its simplicity and speed. The three-stage, semester-wise dynamic prediction process enables university

administrations to monitor a student's risk level throughout the entire course of study and to intervene in a timely manner when necessary.

For future research, it is recommended to test the proposed model on real data from several higher education institutions in Uzbekistan, to practically integrate the model's outputs into tutoring and dean's office services, and to further develop the ethical and legal foundations ensuring the confidentiality of student data. In conclusion, machine-learning-based predictive systems should be regarded not as a means of surveillance but as an early-warning mechanism designed to provide students with timely support and to improve the quality of education.

## REFERENCES

1. Romero, C., Ventura, S. (2010). Educational Data Mining: A Review of the State-of-the-Art. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 40(6), 601–618. https://www.researchgate.net/publication/224160756_Educational_Data_Mining_A_Review_of_the_State_of_the_Art

2. Romero, C., Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery, 10(3), e1355. https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1355

3. Krüger, J. G. C., de S. Britto, A., Barddal, J. P. (2023). An explainable machine learning approach for student dropout prediction. Expert Systems with Applications, 233, 120933. https://doi.org/10.1016/j.eswa.2023.120933

4. Marquez-Vera, C., Morales, C. R., Soto, S. V. (2013). Predicting School Failure and Dropout by Using Data Mining Techniques. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 8(1), 7–14. https://doi.org/10.1109/RITA.2013.2244695

5. Aldowah, H., Al-Samarraie, H., Alzahrani, A. I., Alalwan, N. (2020). Factors affecting student dropout in MOOCs: a cause and effect decision-making model. Journal of Computing in Higher Education, 32(2), 429–454. https://link.springer.com/article/10.1007/s12528-019-09241-y

6. Student Dropout Prediction Using Random Forest and XGBoost Method. INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi, 2025. https://ojs.unpkediri.ac.id/index.php/intensif/article/view/21191

7. Predicting Student Dropout in Higher Education Using Machine Learning Techniques: A Predictive Model Using XGBoost Algorithm. ResearchGate, 2024. https://www.researchgate.net/publication/386564297_Predicting_Student_Dropout_in_Higher_Education_Using_Machine_Learning_Techniques_A_Predictive_Model_Using_XGBoost_Algorithm

8. Machine Learning Models for Predicting Student Dropout, Continued Enrollment and Graduation (Multinomial Logistic Regression, Random Forest, XGBoost). Preprints.org, 2025. https://www.preprints.org/frontend/manuscript/534d2effd7ea0040a8ea024b670f0033/download_pub

9. Predicting student dropouts with machine learning: An empirical study in Finnish higher education. International Journal of Educational Research, ScienceDirect, 2024. https://www.sciencedirect.com/science/article/pii/S0160791X24000228

10. Student dropout prediction through machine learning optimization: insights from Moodle log data. Scientific Reports (Nature), 15, 2025. https://www.nature.com/articles/s41598-025-93918-1

11. SentiDrop: A Multi Modal Machine Learning model for Predicting Dropout in Distance Learning. arXiv preprint, 2025. https://arxiv.org/pdf/2507.10421

12. Chung, J. Y., Lee, S. (2019). Dropout early warning systems for high school students using machine learning. Children and Youth Services Review, 96, 346–353. https://doi.org/10.1016/j.childyouth.2018.11.030

13. Uzbekistan's strategy for the development of artificial intelligence technologies until 2030. Digital Watch Observatory, 2025. https://dig.watch/resource/uzbekinstans-strategy-for-the-development-of-artificial-intelligence-technologies-until-2030

14. The Future of Higher Education in Uzbekistan: The Potential and Challenges of Integrating Artificial Intelligence (National Survey). Perspectives of Higher Education Development, 2025. https://erasmus.uz/journal/index.php/files/article/view/129