



Analysis of Student Academic Performance Through Machine Learning Methods in The Field of Pedagogy

OPEN ACCESS

SUBMITTED 14 March 2025

ACCEPTED 10 April 2025

PUBLISHED 12 May 2025

VOLUME Vol.05 Issue05 2025

COPYRIGHT

© 2025 Original content from this work may be used under the terms of the creative commons attributes 4.0 License.

Ikromov Khusan Kholmakhamatovich

Andijan State Technical Institute, Uzbekistan

Abstract: This article explores the application of machine learning algorithms to analyze, predict, and personalize the educational process based on students' academic performance. Through statistical and computational methods, various student-related attributes were analyzed, with the Random Forest algorithm identified as the most accurate predictive model. The study led to the development of an intelligent system for diagnostic evaluation, personalized approaches, and automated pedagogical recommendations. The findings highlight the significant potential of artificial intelligence tools in enhancing the effectiveness of education.

Keywords: Machine learning, student performance, random forest, pedagogical analysis, artificial intelligence, predictive models, educational statistics, personalized learning, diagnostic system, digital education.

Introduction: In today's society, the advancement of the education system and its integration with digital technologies necessitate the implementation of innovative approaches and tools in pedagogical processes [1]. In higher education institutions, ensuring the quality of educational processes, assessing students' knowledge, and developing them according to their individual learning abilities have become pressing issues. Particularly, conducting in-depth analysis of student knowledge and evaluating their learning performance using precise metrics is of great importance in improving educational quality.

Existing assessment systems are often based on general

approaches, which fail to consider students' individual capabilities and learning pace. Each student may vary in how they perceive, understand, and apply knowledge [2]. As a result, traditional assessment methods are not sufficiently effective in identifying individual differences among students [3]. The incorporation of artificial intelligence and machine learning technologies into education has significantly expanded the ability to analyze student activity [4].

Machine Learning (ML) algorithms enable the identification of students' learning levels, prediction based on prior performance, and development of personalized learning strategies. These technologies hold substantial scientific and technical potential. Especially, analyzing large educational datasets—such as test scores, assignment completion, participation statistics—allows for accurate prediction of learning outcomes, thereby improving educational effectiveness. This approach is not only useful for evaluating students' knowledge but also for identifying their learning challenges and designing personalized pedagogical interventions.

In recent years, considerable research has focused on the implementation of machine learning models in education. Specifically, analyzing data from student activity has yielded measurable improvements in learning efficiency, informed instructor decisions on personalized teaching, and facilitated optimization of curricula and assessment systems. Furthermore, many international universities are actively adopting big data and AI technologies in education, which is transforming pedagogical analysis to a new level.

It is worth noting that Uzbekistan is also undergoing systematic reforms to digitize the education system and integrate information technologies into teaching. The government's "Digital Education" initiative, the introduction of tech-based learning platforms, and the expansion of online and distance learning formats are evidence of these positive developments. However, current educational analysis processes still involve many subjective approaches, limiting objective pedagogical analysis and decision-making.

Therefore, using machine learning methods in pedagogy to automatically analyze student performance and predict learning outcomes presents an opportunity to elevate educational quality to a new level. On one hand, this helps better understand each student's personality and supports a personalized learning approach; on the other hand, it provides educators with crucial information for selecting effective methodological strategies.

The relevance of this research lies in the significant role that innovative technologies—particularly machine

learning models—play in deeply analyzing student performance and developing individualized teaching strategies. Traditional methods often rely on general statistical indicators, while machine learning techniques can identify individual differences, detect critical issues in learning, and offer effective solutions. Additionally, the consistency, reproducibility, and ability to process large datasets make these approaches highly suitable for pedagogical analysis.

The novelty of this study is in its experimental comparison of the effectiveness of several machine learning algorithms in predicting student performance. Modeling was carried out using real-world datasets, and decisions were made based on the most effective model. This allows for pedagogical analysis not only on a statistical basis but in a dynamic, interactive, and predictive manner. Furthermore, the customization of machine learning models to reflect the specific characteristics of the educational process is also addressed.

The primary aim of this research is to explore the potential of using machine learning techniques to analyze, predict, and personalize student learning outcomes. To achieve this goal, the following objectives were defined: identifying and analyzing factors influencing student performance; selecting appropriate machine learning algorithms and testing them on real data; determining the most effective model and developing a prediction system; and assessing the feasibility of implementing personalized learning strategies based on the developed model.

This scientific article discusses the possibilities of applying innovative technologies in the education system to analyze student activity, develop adaptive teaching strategies, and automate pedagogical decision-making. These efforts contribute to modernizing pedagogical processes in line with current demands, enhancing educational quality, and helping each student realize their full potential.

METHODS

This study employed a comprehensive methodological approach aimed at analyzing and predicting students' academic performance using statistical analysis, literature review, and the development of modern machine learning algorithms. Each phase of the research process was purpose-driven, structured through interconnected and complementary methods.

Initially, empirical data required for the study was collected. At this stage, statistical indicators from undergraduate students who participated in the educational process at various higher education institutions in Uzbekistan between 2020 and 2024 were analyzed. Specifically, data included interim

assessments, final exam scores, laboratory work, independent assignments, attendance records, and activity metrics gathered from online learning platforms. These indicators were first prepared for analysis by handling missing values, removing incorrect or outlier entries, recoding necessary fields, and applying normalization. The cleaned dataset was then subjected to statistical analysis.

During the statistical analysis stage, descriptive statistics—such as mean, median, mode, variance, standard deviation, and quartiles—were calculated to gain an overall understanding of the attributes. To test normality, the Kolmogorov-Smirnov and Shapiro-Wilk tests were applied. Pearson and Spearman correlation coefficients were used to explore relationships among attributes. Based on correlation matrices, the influence of each attribute on the final grades was determined, and through factor analysis, the main latent variables were extracted.

Further analysis involved applying analysis of variance (ANOVA) and covariance regression models. These statistical techniques helped to identify and validate both internal and external factors affecting students' learning performance. The findings from this stage informed the selection of attributes for machine learning modeling.

The next phase focused on the analysis of academic literature related to the use of machine learning algorithms in pedagogical contexts. National scholarly articles, master's and doctoral dissertations on higher education in Uzbekistan were reviewed alongside international publications indexed in databases such as Scopus, Web of Science, IEEE Xplore, and Springer. Particular attention was given to models developed and tested in real educational environments, including Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest, K-Nearest Neighbors, Gradient Boosting, and Deep Neural Networks. The analysis revealed that supervised learning algorithms—especially Random Forest and Logistic Regression—demonstrated high accuracy in predicting academic performance [5].

The literature review indicated that pedagogical data is often complex, ambiguous, and multi-dimensional, which limits the effectiveness of classical statistical models. Therefore, machine learning algorithms were justified as optimal tools for uncovering hidden relationships within datasets and improving prediction accuracy. Based on this reasoning, the Random Forest algorithm was selected as the most suitable for this study. This model, which comprises multiple decision trees that make predictions based on small sample evaluations, generates final predictions based on the

majority or average of those trees. It was favored for its stability, resistance to overfitting, and ability to identify feature importance.

To build and evaluate the machine learning model, Python programming libraries were utilized. Data reading, cleaning, and preprocessing were performed using Pandas and NumPy. Visualization, graphics, and correlation matrices were created using Seaborn and Matplotlib. The core model was developed using the Scikit-learn library. The dataset was split into 80% training and 20% testing subsets, and cross-validation (5-fold) was conducted to evaluate the model's generalization capacity.

The model was assessed using key metrics: accuracy, precision, recall, F1-score, and ROC-AUC. Results showed that the Random Forest algorithm could predict students' academic performance with an accuracy of 91–94%. Additionally, the model provided insights into feature importance, helping educators understand which factors most significantly influence student success.

From a scientific standpoint, the primary objective of the machine learning model was not only prediction but also providing actionable recommendations to support pedagogical decision-making. Consequently, a recommendation system was developed based on the model's output. This system enables tailored educational adjustments for low-performing students, increases engagement and participation, promotes independent study, and facilitates personalized support sessions.

The integration of these methods, approaches, and technologies provides a robust theoretical and practical foundation for decision-making in pedagogical activities based on artificial intelligence. The model developed through this research is suitable for real-world educational environments and represents a significant step toward incorporating digital technologies into pedagogical processes.

RESULTS

The primary objective of the machine learning models developed in this study was to create an intelligent system capable of reliably predicting students' academic performance and providing individualized analysis based on their engagement levels within the educational process. The algorithms were tested using statistical data derived from over 5,000 students involved in real-world learning activities. This dataset included variables such as theoretical knowledge scores, laboratory performance, attendance rates, independent work, and assignment completion levels.

Prior to model development, preprocessing was

performed on the attributes. This included ensuring data completeness, normalization, evaluation of attribute variability, and selection based on their direct impact on student performance. Out of 25 initial attributes, 12 were identified as key determinants. The selected algorithms included Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors

(KNN), and Random Forest. Their effectiveness in predicting academic performance was comparatively analyzed.

The evaluation metrics for each model are presented in the following table:

Table 1. Performance Metrics of Selected Machine Learning Algorithms

Algorithm	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	86.5%	84.3%	85.1%	84.7	0.88
Support Vector Machine	88.9%	87.2%	86.5%	86.8	0.90
K-Nearest Neighbors	85.3%	82.1%	83.7%	82.9	0.86
Random Forest	93.2%	92.4%	91.6%	92.0	0.94

As the table indicates, the Random Forest algorithm achieved the highest performance across all evaluation metrics. Notably, its F1-score (92.0) and ROC-AUC value (0.94) demonstrate both high classification accuracy and model stability and sensitivity. This confirms Random Forest as a highly optimized model capable of performing well in uncertain and complex educational environments.

Using this model, students were categorized into "High," "Medium," and "Low" academic performance levels. The Random Forest algorithm correctly classified students with 95% accuracy for the "High" group, 89% for the "Medium" group, and 90% for the "Low" group. The results were analyzed using a confusion matrix, and diagnostic evaluations were conducted on misclassified cases.

Feature importance analysis was also performed. Indicators such as midterm scores, laboratory performance, and timely assignment completion were found to significantly impact overall student performance. The feature importance graph generated by the Random Forest model confirmed these findings, providing educators with actionable insights for effective pedagogical strategies.

The model also enables real-time monitoring of student progress and the delivery of personalized recommendations. For instance, low-performing students can automatically receive alert messages, additional assignments, or individualized consultation offers. Such functionality creates valuable opportunities to implement AI-based support within educational practice.

The findings indicate that diagnostic models built with machine learning algorithms can effectively guide individualized student development, initiate early warning systems, and support personalized learning processes. These results provide a solid scientific and practical foundation for the digital transformation of

pedagogical management systems.

DISCUSSION

The results of the study clearly demonstrate that among the machine learning algorithms used to assess and predict student performance, the Random Forest model achieved the highest level of effectiveness. It outperformed other models across all major evaluation metrics—accuracy, recall, precision, F1-score, and ROC-AUC. This underscores the model's capacity to efficiently operate on complex, multidimensional, and interrelated pedagogical data.

A key strength of Random Forest lies in its ensemble mechanism, which aggregates predictions from multiple decision trees and bases final outputs on their average or majority vote. This structure mitigates the risk of overfitting and enhances the model's generalizability. Such an approach is particularly important when analyzing multifaceted data reflecting various aspects of student behavior—such as attendance rates, laboratory performance, midterm exams, and timely task completion. Additionally, the model demonstrated strong handling of uncertainty and ambiguity across input attributes [6].

Compared to other tested models, Random Forest uniquely provides insights into the influence of each attribute on a student's overall academic performance. This feature not only enhances predictive accuracy but also supports diagnostic evaluation. For example, the most impactful attributes—laboratory results, independent assignments, and attendance—clearly illustrate which elements contribute most to academic success. These insights can guide educators and academic advisors in refining focus areas within the learning process.

Another advantage of the developed platform is its ability to monitor students' individual learning trajectories in real time and to detect at-risk learners at early stages. The findings show that many low-

performing students consistently lagged across specific attributes; however, with predictive analytics and early alert systems, such issues can be addressed proactively. The Random Forest model exhibited very high sensitivity in identifying “at-risk” students, making it not only scientifically robust but also highly relevant for practical educational implementation.

The model’s robustness to diverse and dynamic data sources further enhances its applicability. Despite variations in the dataset, the model consistently maintained reliable performance. This indicates strong potential for adapting the system across various higher education institutions. There is even the possibility of scaling the model for nationwide integration into the educational system.

CONCLUSION

In conclusion, the application of the Random Forest algorithm in the educational field offers more than just statistical predictions. It enables the personalization of learning, identification of individual needs, and enhancement of teaching practices—delivering innovative solutions aligned with the goals of modern digital education. This supports evidence-based pedagogical decision-making through the use of machine learning methods.

REFERENCES

Ikromov, X. TA’LIM JARAYONIDAGI SUN’IY INTELLEKTGA ASOSLANGAN AXBOROTNI QAYTA ISHLASH VOSITALARINING TAHLILI. O ‘ZBEKISTON RESPUBLIKASI OLIY TA’LIM, FAN VA INNOVATSIYALAR VAZIRLIGI ANDIJON DAVLAT UNIVERSITETI UMUMIY PEDAGOGIKA KAFEDRASI, 315.

Umidjon’s, K. I., & Ilhomjon’s, S. D. (2024). EFFECTIVENESS OF BUSINESS PROCESS AUTOMATION IN GROCERY STORES. *International Journal of Advance Scientific Research*, 5(12), 279-284.

IKROMOV, X. (2024). TALABALARNI MA’LUMOTLAR BAZASINI BOSHQARISH ASOSIDA INNOVATSION AXBOROT TIZIMLARINI ISHLAB CHIQISHGA O ‘RGATISH METODIKASI. *News of the NUUZ*, 1(1.1. 1), 93-96.

Baker, R. S., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. In *Learning Analytics* (pp. 61–75). Springer.

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.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.