



<http://dx.doi.org/10.35596/1729-7648-2026-32-3-43-51>

UDC 004.85

MACHINE LEARNING-BASED PREDICTION OF STUDENTS' ACADEMIC PERFORMANCE IN HIGHER EDUCATION

XUNHUAN REN

Belarusian State University of Informatics and Radioelectronics (Minsk, Republic of Belarus)

Abstract. In higher education, student academic performance is considered a key educational outcome. Consequently, improving academic performance is a pressing objective for many higher education institutions. Predicting and classifying student performance can provide valuable insights and facilitate the development of appropriate strategies to enhance academic results. In recent years, data mining methods have been widely applied and have demonstrated significant success in classification and prediction tasks. This article presents and analyzes the EASP student performance dataset. Seven machine learning methods were applied to classify and predict student academic performance: Multinomial Naïve Bayes, Support Vector Machines, Decision Trees, Logistic Regression, Random Forest, Gradient Boosting, and k-Nearest Neighbors. The performance of these methods was evaluated using precision, recall, the area under the ROC curve (AUC), and the F1-score. Experimental results demonstrated that Logistic Regression achieved the best predictive performance on the EASP test set, with a precision of 0.75.

Keywords: higher education, academic performance, student performance prediction, educational data mining, machine learning.

Conflict of interests. The author declares that there is no conflict of interests.

For citation. Ren X. (2026) Machine Learning-Based Prediction of Students' Academic Performance in Higher Education. *Digital Transformation*. 32 (3), 43–51. <http://dx.doi.org/10.35596/1729-7648-2026-32-3-43-51>.

ПРОГНОЗИРОВАНИЕ АКАДЕМИЧЕСКОЙ УСПЕВАЕМОСТИ СТУДЕНТОВ В ВЫСШЕМ ОБРАЗОВАНИИ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

С. ЖЭНЬ

*Белорусский государственный университет информатики и радиоэлектроники
(Минск, Республика Беларусь)*

Аннотация. В высшем образовании академическая успеваемость студентов рассматривается как важный образовательный результат. Поэтому повышение академической успеваемости является актуальной задачей для многих учреждений высшего образования. Прогнозирование и классификация успеваемости студентов могут предоставить ценную информацию и способствовать разработке соответствующих стратегий для улучшения академических результатов. В последние годы методы интеллектуального анализа данных широко применяются и демонстрируют значительные успехи в задачах классификации и прогнозирования. В статье представлен и проанализирован набор данных об успеваемости студентов EASP. Для классификации и прогнозирования академической успеваемости студентов применены семь методов машинного обучения: мультиномиальный наивный байесовский классификатор, метод опорных векторов, дерево решений, логистическая регрессия, случайный лес, градиентный бустинг и метод *k*-ближайших соседей. Эффективность этих методов оценивалась с использованием прецизионности, полноты, площади под ROC-кривой (AUC) и F1-меры. Экспериментальные результаты показали, что логистическая регрессия достигает наилучшей прогнозной эффективности на тестовой выборке EASP с прецизионностью 0,75.

Ключевые слова: высшее образование, академическая успеваемость, прогнозирование успеваемости студентов, интеллектуальный анализ образовательных данных, машинное обучение.

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Для цитирования. Жэнь, С. Прогнозирование академической успеваемости студентов в высшем образовании на основе методов машинного обучения / С. Жэнь // Цифровая трансформация. 2026. Т. 32, № 3. С. 43–51. <http://dx.doi.org/10.35596/1729-7648-2026-32-3-43-51>.

Introduction

In the post-COVID era, higher education institutions (HEIs) have increasingly considered educational reforms aimed at improving the quality of higher education for future generations [1]. The quality of higher education can be assessed using students' grade point average (GPA), as GPA is commonly regarded as an important indicator of academic performance [2]. Previous studies have also shown that GPA is closely associated with university graduates' future earnings [3]. Therefore, improving students' academic performance, particularly their GPA, has become an important objective of higher education reform.

Previous studies have investigated various factors associated with students' GPA, including college entrance examination scores [4], alcohol and marijuana use [5], relationships with family members and friends [6], race and gender [7], and the amount of time spent studying and working [8]. These studies provide useful insights into the factors influencing academic performance. However, most of them focus on individual factors, whereas improving students' academic performance generally requires the simultaneous consideration of multiple dimensions.

Data mining has attracted considerable attention in recent years. By applying machine learning or deep learning methods, it can efficiently extract valuable information from high-dimensional feature spaces for classification and prediction. In the field of higher education, the application of data mining has gradually developed into educational data mining (EDM), which has been widely investigated for student performance prediction and educational decision support [9–11].

According to the types of datasets used, existing EDM studies on student performance prediction can generally be divided into studies based on open-source datasets and those based on private datasets. One commonly used open-source dataset is the UCI Student Performance dataset, which contains a range of student-related features and has been adopted in numerous previous studies [12]. However, the data were collected before the COVID-19 pandemic and may not fully reflect the current context of higher education. Other studies have used private datasets, which are often more recent and contain a wide range of student-related features [13, 14]. Nevertheless, such datasets are usually difficult to access, which limits their reuse, reproducibility, and comparative analysis.

The effects of alcohol on student performance (EASP) dataset [15] is a recently released open-source dataset containing information collected in 2023 from students at Stellenbosch University in South Africa. The dataset includes a relatively large number of samples and a wide range of relevant features, such as college entrance examination scores, alcohol consumption, family relationships, gender, time spent studying and working, and other student-related variables. Therefore, it provides a useful platform for educational data mining research.

However, to the best of our knowledge, few studies have used the EASP dataset to predict students' academic performance. Therefore, this study applies data mining techniques to predict student performance using this dataset. To simplify the prediction task, student performance is formulated as a binary classification problem. Students whose GPA is higher than the mean GPA are assigned to one class, whereas the remaining students are assigned to the other class.

The remainder of this paper is organized as follows. Section 2 describes the EASP dataset and the corresponding data preprocessing procedures. Section 3 presents the overall methodology. Section 4 reports and discusses the experimental results. Finally, Section 5 concludes the paper.

EASP dataset description

The EASP dataset [15] was constructed from a behavioral survey covering factors related to students' daily lives, study habits, and academic performance. These variables may provide useful information for predicting students' academic outcomes and may support the timely adjustment of study strategies.

The dataset contains 406 samples, each described by 15 independent features. However, several variables contain missing values. Therefore, appropriate imputation strategies are applied during data preprocessing. Missing values in the gender variable are imputed by randomly assigning either male or female. Missing values in the 2023 academic-year average GPA are estimated based on the diffe-

rence between a student's admission average score and the overall mean admission score, together with the overall mean GPA for the 2023 academic year. Fig. 1 shows the distribution of students' average GPA in the 2023 academic year.

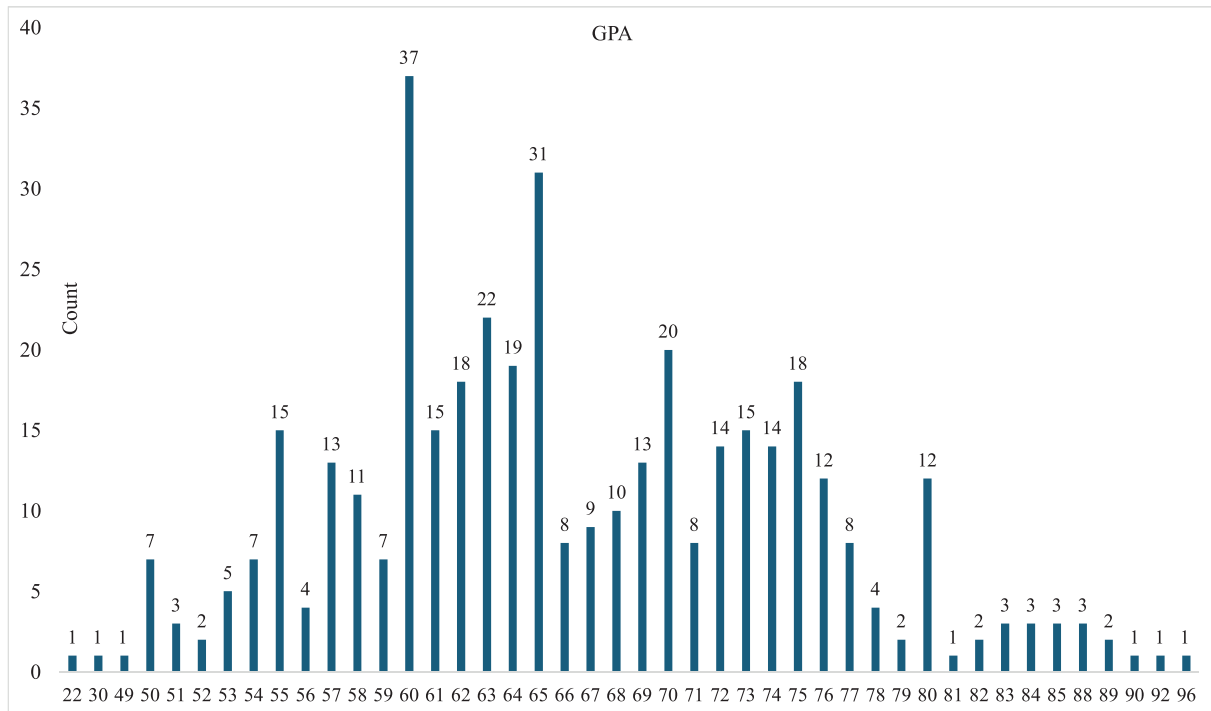


Fig. 1. Students' academic year average grade point average in 2023 academic year

We first calculated the mean and standard deviation of the students' average GPA for the 2023 academic year, which were 66.36 and 9.00, respectively. The students were then divided into two groups according to whether their average GPA was lower than 66.36. Students with an average GPA lower than 66.36 were assigned to one class, whereas the remaining students were assigned to the other class. Tab. 1 presents all 15 features and their corresponding descriptions. In addition, Fig. 2 shows the distributions of six representative features for visualization purposes: gender, additional study hours per week, number of missed classes per week, alcohol consumption, number of failed modules, and frequency of socializing.

Table 1. Features and text descriptions

Feature	Type	Description
Sex	Categorical	Sex or gender of the student
Matric Average	Numerical, continuous	The student's Grade 12 matriculation average, expressed as a percentage
Study Year	Categorical	The student's current year of study
Faculty	Categorical	The faculty at Stellenbosch University to which the student's degree program belongs
Accommodation Status	Binary	Whether the student lives in university accommodation or private accommodation
Monthly Allowance	Categorical	The amount of monthly allowance received by the student while studying in Stellenbosch
Scholarship	Binary	Whether the student receives financial aid or a scholarship
Additional Study Hours per Week	Numerical, discrete	The number of hours spent studying each week in addition to attending scheduled lectures
Socializing Frequency	Numerical, discrete	The frequency with which the student goes out with friends for activities such as clubbing, drinking, or dining
Alcohol Consumption	Numerical, discrete	The number of alcoholic drinks consumed when the student goes out

Ending of Tab. 1

Feature	Type	Description
Missed Classes per Week	Numerical, discrete	The number of classes missed by the student each week
Failed Modules	Numerical, discrete	The number of academic modules the student has failed to date
Romantic Relationship Status	Binary	Whether the student is currently in a romantic relationship
Parental Approval of Alcohol Consumption	Binary	Whether the student's parents approve of the student consuming alcohol
Relationship with Parents	Categorical	The degree of closeness between the student and their parents

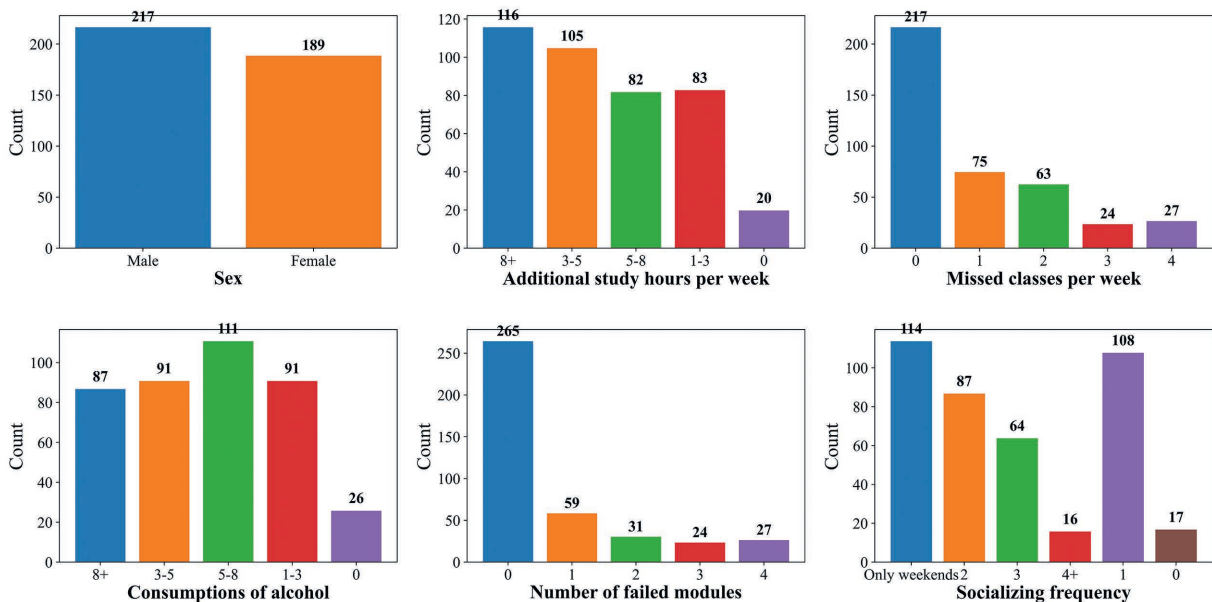


Fig. 2. Distributions of six representative features in the effects of alcohol on student performance dataset

Methodology

The proposed data mining methodology follows a conventional machine learning pipeline. Seven classification algorithms are employed: support vector machine, decision tree, logistic regression, random forest, k -nearest neighbors, multinomial naive Bayes, and gradient boosting. To ensure an objective performance evaluation, the original dataset is divided into two subsets: a training set and an independent test set. The training set is used for model training, hyperparameter optimization, and cross-validation, whereas the test set is used only for the final evaluation of the trained models.

Grid search combined with five-fold cross-validation is applied to determine the optimal hyperparameter configuration for each classifier. The overall structure of the proposed data mining pipeline is illustrated in Fig. 3.

Grid search is a systematic hyperparameter optimization method that evaluates a predefined set of hyperparameter combinations. For each combination, model performance is estimated using five-fold cross-validation, and the configuration with the best validation performance is selected. Although this procedure increases computational cost, it provides a more reliable basis for hyperparameter selection than evaluation based on a single training-validation split.

Cross-validation is an important technique for evaluating the generalization ability of machine learning models. In five-fold cross-validation, the training set is divided into five approximately equal subsets. In each iteration, four subsets are used for training and the remaining subset is used for validation. This process is repeated five times so that each subset is used once as the validation set. The average validation performance is then used for model selection and hyperparameter tuning, thereby reducing the risk of overfitting to a particular data split.

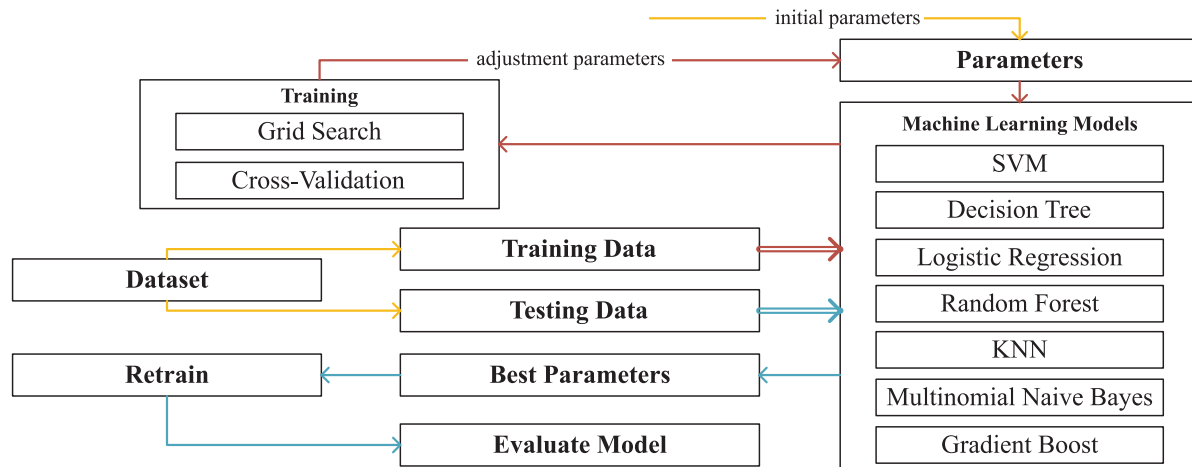


Fig. 3. Overall structure of the data mining pipeline

A. Description of the classifiers

As mentioned above, seven different classifiers are included in the proposed pipeline. Before evaluating their performance on the EASP dataset, it is necessary to briefly review the basic principles of these classifiers and their main hyperparameters.

Support vector machine (SVM) is a classical supervised learning algorithm widely used for classification, regression, and anomaly detection. The basic principle of SVM is to find an optimal separating hyperplane in the feature space so that training samples from different categories can be separated as clearly as possible. This improves the generalization ability of the model and reduces the risk of overfitting.

Decision tree (DT) is a simple and efficient supervised machine learning model that has important applications in many scenarios, such as classification and prediction, feature engineering, anomaly detection, and rule extraction. It is a non-parametric learning algorithm based on a tree structure that recursively divides samples into different child nodes and ultimately produces an interpretable prediction model. When constructing a decision tree, an optimal feature-splitting criterion needs to be selected. Such criteria, including information gain and the Gini index, measure the effectiveness of features in dividing the samples, allowing the features that have the greatest impact on the results to be automatically selected.

Logistic regression (LR) is a supervised learning model widely used for classification problems. As a probabilistic model, LR uses an S-shaped sigmoid function to predict the class probabilities of samples, thereby performing binary or multiclass classification. The LR model is simple and easy to understand, has few parameters, and can be trained rapidly. Therefore, it can effectively handle linearly separable binary classification problems.

Random forest (RF) is an ensemble learning algorithm based on decision trees. It improves prediction accuracy by combining multiple decision tree models and can effectively address the problem of overfitting. Random forest is widely used in many machine learning tasks, such as regression, classification, and clustering, and is favored for its simplicity and robustness.

The k -nearest neighbours (KNN) algorithm is an instance-based, non-parametric learning algorithm that is simple and intuitive and performs well in a variety of machine learning tasks, such as classification, regression, and clustering. It is also known as a “lazy learning” algorithm because it identifies the (K) training samples most similar to a new input instance and predicts the label or value of the new instance based on their class labels or values. The number of neighbours is the most important hyperparameter of the KNN algorithm, as it determines the number of neighbouring training samples considered when predicting new samples. When (K) is small, KNN tends to produce high variance, potentially overfitting the training set. When (K) is large, KNN may produce high bias, potentially resulting in underfitting and failing to capture the complex features of the data.

Naive Bayes (NB) is a simple and efficient classification algorithm based on Bayes’ theorem and is widely used in machine learning and data mining. It is based on the assumption of conditional

independence and achieves fast classification of unknown samples by calculating the conditional probabilities between sample features and classes. It has demonstrated excellent performance in many practical applications, such as spam filtering and text classification. Multinomial naive Bayes is a variant of the Naive Bayes algorithm specifically designed to deal with problems involving discrete features, such as text classification.

Gradient boosting (GB) is an ensemble learning algorithm based on an additive model. It iteratively constructs weak learners, such as decision trees, and optimizes the loss function using gradient descent to obtain a robust prediction model. Unlike parallel ensemble methods such as random forest, gradient boosting uses a sequential ensemble approach to improve the model's fit to the training data in each iteration. The gradient boosting algorithm has been widely used in several machine learning fields, such as regression and classification, and has achieved excellent performance.

B. Model hyperparameter tuning using grid search

We optimized the hyperparameters using a grid search technique to identify the parameter combinations that produced the best algorithmic performance. Grid search is a widely used hyperparameter optimization technique that first defines a predefined grid of hyperparameter values and then evaluates each parameter combination using cross-validation. The parameter setting that produces the best results is ultimately selected. Through this systematic parameter search and performance evaluation, the potential of each algorithm can be fully utilized, and its performance in practical applications can be improved. The detailed process of hyperparameter optimization using grid search and cross-validation is described as follows:

- 1) select a specific algorithmic model, such as DT or RF;
- 2) determine the hyperparameters to be optimized and their value ranges and combine them into a parameter grid;
- 3) traverse all hyperparameter combinations using grid search;
- 4) evaluate the average performance of each combination using cross-validation. In this study, five-fold cross-validation is used;
- 5) select the parameter combination with the highest cross-validation score.

The combination of grid search and cross-validation can effectively reduce the risk of overfitting and identify the optimal hyperparameter settings within the predefined search space. It systematically explores the parameter space and reduces the influence of randomness in sample selection, making it a commonly used and reliable optimization method.

Experiments and result discussion

The main metrics used to evaluate the performance of the model include precision, recall, F1-score and AUC. Their calculation procedure is shown below:

$$\text{Precision} = \frac{TP}{TP + FP}; \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN}; \quad (2)$$

$$\text{F1-Score} = \frac{2 \cdot (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}, \quad (3)$$

where TP (true positive) is number of correctly predicted positive samples; FP (false positive) is number of incorrectly predicted positive samples; FN (false negative) is number of samples that are incorrectly predicted to be negative.

$$\text{AUC} = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} I(S(Pos(x_i)) > S(Neg(x_j))), \quad (4)$$

where n_+ is the number of positive samples; n_- is the number of negative samples; $S(Pos(x_i))$ is the predicted score for the i -th positive sample; $S(Neg(x_j))$ is the predicted score for the j -th negative sample; $I(\cdot)$ is an indicator function that returns 1 if the condition is true and 0 otherwise.

The AUC value ranges from 0 to 1. A higher AUC value indicates better model performance. The experimental results obtained using different classifiers on the EASP dataset are shown in Tab. 2.

Table 2. Experimental results obtained using different classifiers

Algorithm	Precision on the training set	Best parameters of the model	Test precision	Test recall	Test AUC	Test F1-score
SVM	0.77	C = 10, Linear Kernel	0.73	0.64	0.78	0.67
DT	0.80	Max Depth = 3, Min Samples Split = 2	0.75	0.64	0.77	0.69
LR	0.74	C = 100, Penalty = 12	0.75	0.68	0.79	0.70
RF	0.81	Max Depth = 3, Max Features = sqrt, Min Samples Split = 2	0.75	0.58	0.80	0.67
KNN	0.72	N Neighbors = 7	0.63	0.55	0.70	0.56
Multinomial NB	0.66	Alpha = 0.01, Fit Prior = True	0.68	0.77	0.73	0.68
GB	0.82	N Estimators = 50, Learning Rate = 0.1, Max Depth = 2	0.75	0.64	0.81	0.69

From Tab. 2, it can be seen that the logistic regression algorithm achieves a precision of 0.74 on the training set. With the best model parameters on the test set precision is 0.75, recall is 0.68, AUC is 0.79 and F1-score is 0.7. Compared to other algorithms logistic regression algorithm has the best overall performance on the test set. The gradient boosting algorithm achieves a precision of 0.82 on the training set, which is the highest of these algorithms. With the best parametric model on the test set precision is 0.75, recall is 0.64, AUC is 0.81 and F1-score is 0.69. It is slightly inferior to the logistic regression algorithm in terms of overall performance on the test set, but superior to the other algorithms.

Fig. 4 and 5 show the confusion matrices and ROC curves of all seven classifiers on the test set, respectively.

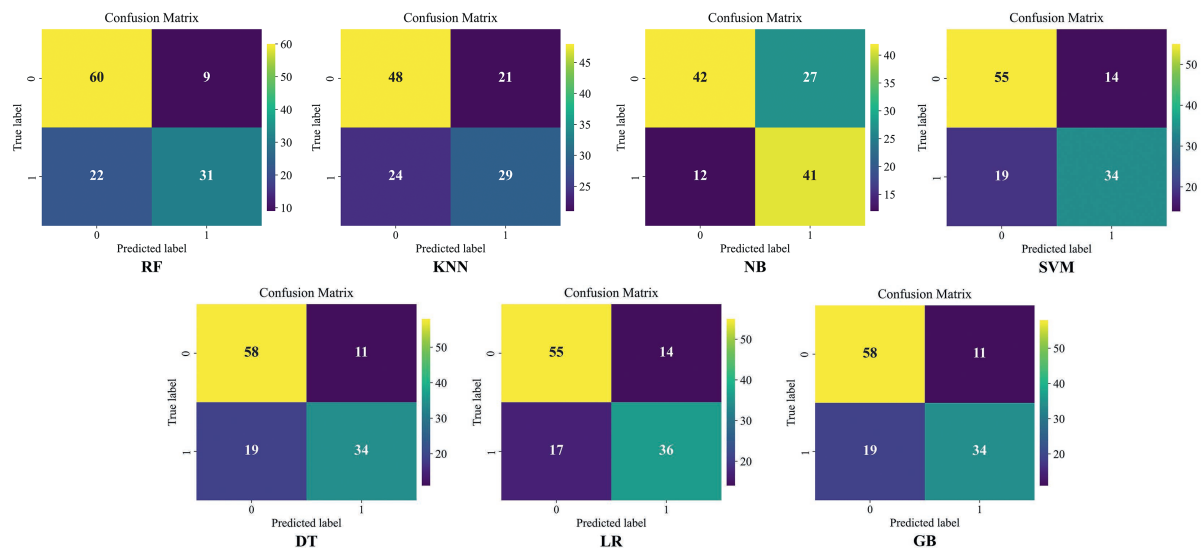


Fig. 4. Confusion matrices of the seven classifiers on the test set

Conclusion

1. In this paper, student academic performance prediction was investigated using the recently released effects of alcohol on student performance dataset. Seven widely used machine learning classifiers were evaluated, including support vector machine, decision tree, logistic regression, random forest, *k*-nearest neighbours, multinomial naive Bayes, and gradient boosting. Grid search and five-fold cross-validation were used to optimize the hyperparameters of each model.

2. The experimental results show that gradient boosting achieved the highest training precision of 0.82. However, logistic regression demonstrated the best overall performance on the independent test set, achieving a precision of 0.75, recall of 0.68, AUC of 0.79, and F1-score of 0.70. These results indicate that relatively simple machine learning models can still provide competitive and robust performance for student academic performance prediction when the dataset size is limited.

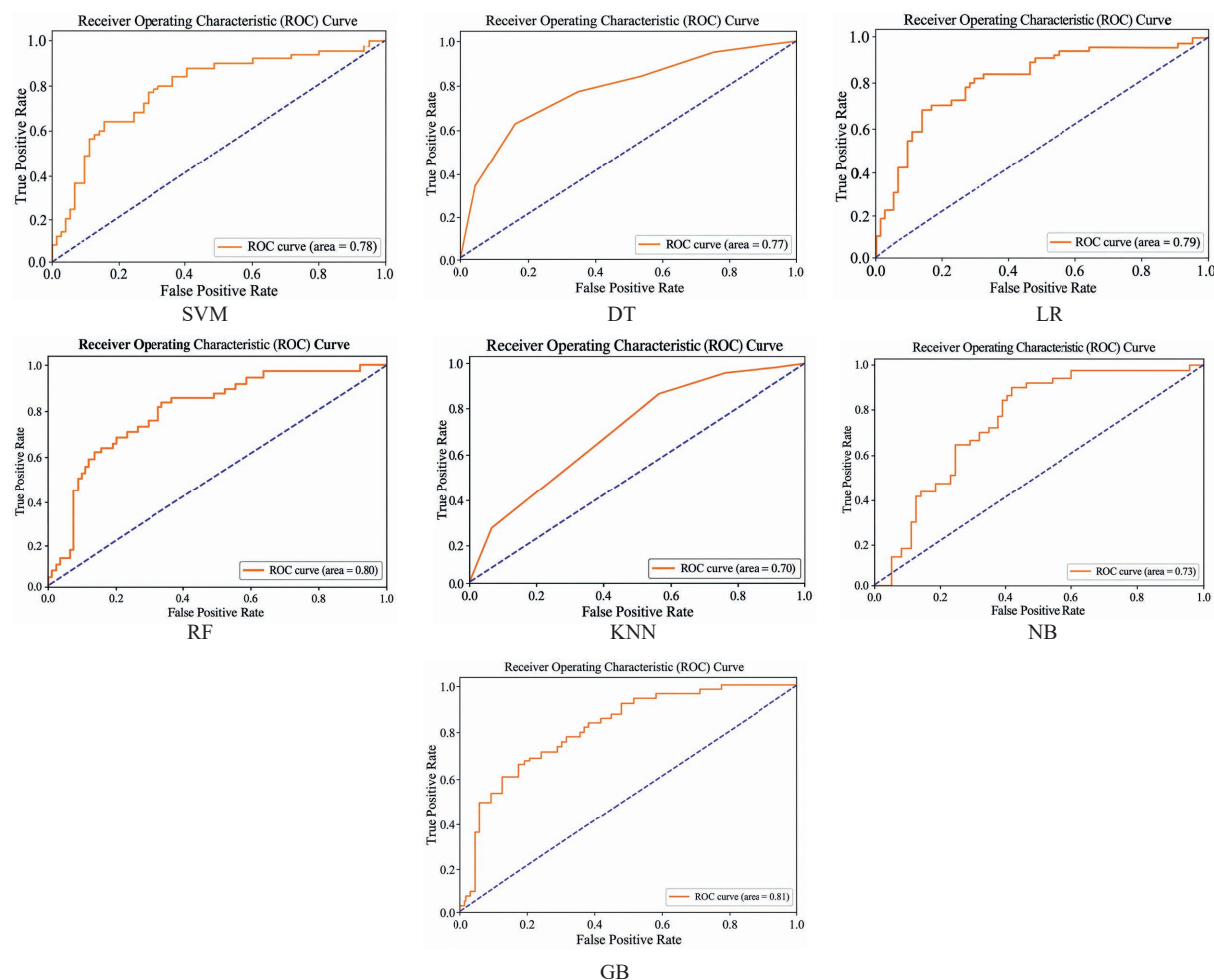


Fig. 5. ROC curves of the seven classifiers on the test set

3. From a practical perspective, the proposed machine learning framework can be adapted for use in higher education institutions in the Republic of Belarus, including the Belarusian State University of Informatics and Radioelectronics (BSUIR). By using locally collected student data, such as academic records, attendance information, study workload, and other relevant educational indicators, the framework could be used to identify students at risk of low academic performance at an early stage. The obtained predictions could support teachers, academic advisers, and university administrators in providing timely academic assistance and developing more targeted educational interventions. At BSUIR, such an approach could potentially be integrated into educational analytics and student monitoring systems to support data-driven academic management. However, before practical deployment, the models should be retrained and validated using data collected from BSUIR students in order to account for differences in the educational environment, student population, and institutional characteristics.

4. Future work will focus on collecting and analyzing local student data from Belarusian higher education institutions, particularly BSUIR, expanding the dataset, improving feature engineering, and exploring more advanced ensemble learning and deep learning methods for educational data mining.

References

1. Rasli A., Fei Z., Abas I. H., Tee M., Prestianawati S. A., Lajuma S., et al. (2024) A Framework for Higher Education Institutions Sustainability: A Multi-Method Study. *Journal of Infrastructure, Policy and Development*. 8 (5). <https://doi.org/10.24294/jipd.v8i5.3566>.
2. Daka H., Changwe R. (2020) An Exploration of Education Quality in the Light of the Grade Point Average and Examination Attrition Rate. *International Journal of Humanities Social Sciences and Education*. 7 (6), 196–207. <http://dx.doi.org/10.20431/2349-0381.0706021>.

3. Zou T., Zhang Y., Zhou B. (2022) Does GPA Matter for University Graduates' Wages? *New Evidence Revisited. PLoS One.* 17 (4). <https://doi.org/10.1371/journal.pone.0266981>.
4. Bai C. E., Chi W., Qian X. (2014) Do College Entrance Examination Scores Predict Undergraduate GPAs? A Tale of Two Universities. *China Economic Review.* 30, 632–647.
5. Meda S. A., Gueorguieva R. V., Pittman B., Rosen R. R., Aslanzadeh F., Tennen H., et al. (2017) Longitudinal Influence of Alcohol and Marijuana Use on Academic Performance in College Students. *PLoS One.* 12 (3). <https://doi.org/10.1371/journal.pone.0172213>.
6. Wider W., Tanucan J. C. M., Lin J., Jiang L., Udang L. N. (2023) Who Makes a Better University Adjustment Wingman: Parents or Friends? *PLoS One.* 18 (12). <https://doi.org/10.1371/journal.pone.0294658>.
7. Bottia M. C., Giersch J., Mickelson R. A., Stearns E., Moller S. (2015) Distributive Justice Antecedents of Race and Gender Disparities in First-Year College Performance. *Social Justice Research.* 29 (1), 35–72.
8. Nonis S. A., Hudson G. I. (2006) Academic Performance of College Students: Influence of Time Spent Studying and Working. *Journal of Education for Business.* 81 (3), 151–159.
9. Albreiki B., Zaki N., Alashwal H. (2021) A Systematic Literature Review of Students' Performance Prediction Using Machine Learning Techniques. *Education Sciences.* 11 (9). <https://doi.org/10.3390/educsci11090552>.
10. Yağcı M. (2022) Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms. *Smart Learning Environments.* 9. <https://doi.org/10.1186/s40561-022-00192-z>.
11. Dutt A., Ismail M. A., Herawan T. (2017) A Systematic Review on Educational Data Mining. *IEEE Access.* 5, 15991–16005. <https://doi.org/10.1109/ACCESS.2017.2654247>.
12. Mansur B. F., Yusof N. (2018) The Latent of Student Learning Analytic with K-Mean Clustering for Student Behaviour Classification. *Journal of Information Systems Engineering and Business Intelligence.* 4 (2), 156–161. <https://doi.org/10.20473/JISEBI.4.2.156-161>.
13. Hasan H. M. R., Rabby A. S. A., Islam M. T., Hossain S. A. (2019) Machine Learning Algorithm for Student's Performance Prediction. *2019 10th International Conference on Computing, Communication and Networking Technologies.* 1–7. <https://doi.org/10.1109/ICCCNT45670.2019.8944629>.
14. Salah H. A., Akeel A. W., Khalaf H. A. (2020) Student Performance Prediction Model Based on Supervised Machine Learning Algorithms. *IOP Conference Series: Materials Science and Engineering.* 928 (3). <https://doi.org/10.1088/1757-899X/928/3/032019>.
15. *Effects of Alcohol on Student Performance Dataset.* Available: <https://www.kaggle.com/datasets/joshuanaude/effects-of-alcohol-on-student-performance>.

Received: 16 July 2026

Accepted: 2 September 2026

Available on the website: 12 October 2026

Information about the author

Ren X., Cand. Sci. (Tech.), Associate Professor at the Department of Infocommunication Technologies, Belarusian State University of Informatics and Radioelectronics

Address for correspondence

220013, Republic of Belarus,
Minsk, P. Brovki St., 6
Belarusian State University of Informatics and Radioelectronics
Tel.: +375 25 605-85-95
E-mail: renxunhuan@bsuir.by
Ren Xunhuan