Journal of Research and Advances in Mathematics Education

Volume 9, Issue 4, October 2024, pp. 190-204
DOI: 10.23917/jramathedu.v9i4.4643
p-ISSN: 2503-3697, e-ISSN: 2541-2590

Predictive analytics of student performance: multi-method and code

Alla Yu. Vladova[1], Katsiaryna M. Borchyk[2]

[1] *Financial University under the Government of the Russian Federation, Moscow, Russia*
[2] *Belarusian-Russian University, Mogilev, Belarus*

Citation: Vladova, A., & Borchyk, K. M. (2024). Predictive analytics of student performance: Multi-method and code. *JRAMathEdu (Journal of Research and Advances in Mathematics Education)*, *9*(4), 190-204. https://doi.org/10.23917/jramathedu.v9i4.4643

ARTICLE HISTORY:
Received 28 March 2024
Revised 16 October 2024
Accepted 23 October 2024
Published 30 October 2024

KEYWORDS:
Academic performance
Educational data analysis
Customized learning

ABSTRACT

The maintenance of a high level of education in universities can be a challenging task due to low academic performance. Despite the significant amount of collected diagnostic data, education managers underutilize machine learning methods to improve the accuracy of predicting academic performance. Authors apply a multi-method approach for data analysis using simple logistic and linear regressions, k-means clustering, that all together gave a synergetic effect. The proposed approach differs from known analogs in that, firstly, the dimensionality of the feature space increases due to the normalization of scores onto a single scale and the creation of new features: the index and rank of students, as well as the changes in performance across various activities for each student. Secondly, students at academic risk are forecasted, and the statistical significance of the features included in the model is evaluated. Thirdly, for each student, the final score for the semester is forecasted using an linear regressive model of academic performance. Fourthly, groups of students with similar learning trajectories are identified for customization of consultations. The authors managed to achieve a high predictive ability of models based on historical training data: binary prediction of exam passing in 90% of cases, prediction of individual assessment in 70% of cases.

INTRODUCTION

The academic performance of students is one of the most important characteristics of the educational activities of an educational institution, by which professors and education managers can judge the results achieved or the problems that exist. Each university has its own systems for assessing academic performance, including various indicators of academic activities (Elisabeta & Alexandru, 2018). The academic performance of students in mathematical disciplines is usually assessed through computer tests, expert evaluation of semester projects, preparation level for seminars, and attendance (Zafar et al., 2020). The quality of students' work is then used for effective educational process management, in making decisions about awarding state academic and named scholarships, issuing diplomas with honors, and other tasks. Thus, the research covers the following tasks: (1) how to effectively use historical data to predict student performance; (2) predictive models that are the most understandable to education managers; (3) how to reduce the subjective influence of experts on the final grades of students.

Literature review

The researchers (Zhang et al., 2021) state that predicting student performance helps all stakeholders in the educational process. For example, students can choose appropriate courses or exercises and make plans for the semester accordingly (Ibrahim & Rusli, 2007), discovering the

*Corresponding author: alla.vladova@gmail.com

relationships between courses. Professors can adjust educational materials and curricula depending on students' abilities and identify students at risk within the group (Kloft et al., 2014). Managers in the education sector can review the curriculum and optimize the set of disciplines. The prediction could guide course selection and early warning on student learning, but finding the key factors affecting most education behaviors is a more important task. That is because (1) the key feature could correspond to interventions of education; (2) the reason of success or failure could reflect the pattern of student learning; (3) understanding of these factors could provide plan settings, course assignments, and learning sequence with suggestions. As (Kahramanoğlu, 2018) notes, the same characteristics help to indirectly analyze the hard and soft skills of prospective teachers.

The article (Yağcı, 2022) proposes a machine learning model for predicting the final scores of undergraduate students, using their scores for midterm exams as the input data. To forecast the exam scores, the performance metrics of random forest, k-nearest neighbor, support vector machines, logistic regression, and naive Bayes algorithms are calculated and compared. The dataset consisted of the academic performance scores of 1854 students at a state university in Turkey during the autumn semester of 2019-2020. Predictions are made using three types of features: midterm exam scores, as well as department and faculty names. The proposed model achieved a classification accuracy of 70–75%. The insufficient accuracy of the model can be explained by the presence of low-variable features.

Authors of the study (Oluwadele et al., 2023) assessed academic performance in the field of medical education through indicators of students' acquisition and perception of knowledge, level of confidence, ease of use of the e-learning platform and willingness to recommend e-learning. The flaw of the proposed approach lies in the qualitative nature of the analyzed features, their strong dependence on the opinion of different experts.

Researchers (Liu & Yu, 2023) uses the online student actions that the e-learning platform allows you to collect, namely: the time it takes to answer a question or submit an assignment, the number of missed questions, excessive tardiness, cheating on tests, derogatory comments in online discussions. The disadvantage of this approach is that the data was taken without additional transformations that affect the performance of the model.

Exploration (Ye et al., 2022) offers a model for predicting the effectiveness of online learning, based on the selection and merging of features. The model uses the relationship between behaviors and examines whether combinations of behaviors are better predictors of academic performance.

The researchers (Yadav & Deshmukh, 2023) emphasize that the most significant factors influencing students' academic performance are low initial scores, family support, living arrangements, gender, previous performance, students' internal assessment, average academic performance, and students' activity in e-learning. They also note that the plan to improve students' academic performance should take into account additional consultations for students with low performance. This helps both students and teachers to overcome the challenges faced during education. The idea of selecting students for additional consultations formes the basis of the fourth stage of the proposed method.

Thus, it is possible to identify the following features that are used to predict academic performance: attendance coefficient; ratio of scores for work or campus activities to the total possible certification score; performance dynamics, the change in scores between the first and second certification. This change in estimates is the basis of the proposed method.

The authors (Yang et al., 2018) designed several week learning activity, includes homework, quizzes, video-based learning. They applied multiple linear regression model to predict students' academic scores. They also reworked the well-known metrics for assessing the accuracy of the models, using data obtained during the cross-validation stage. They believe, however, that the models are applicable to the courses, learning activities and data attributes for which they were developed.

Authors of the article (Arzamastsev et al., 2018) considered various classification methods, including decision trees, naïve Bayesian classifier, random forest, artificial neural networks, linear discriminant analysis, support vector method and their ensembles. The effectiveness of the classifiers is estimated using indicators such as the area under the AUC curve, accuracy, F1 measure and

Figure 1. Tag cloud of scores

classification time. The authors point out that for the four studied datasets (after applying the principal component method to reduce the feature space), virtually identical estimates of F1, accuracy, and AUC were obtained. Authors (Urrutia-Aguilar et al., 2016) used logistic regression to identify medical students who are at academic risk. They binarized the outcome variable that was academic performance and it was coded as "0" if the student had failed at least one subject of the biomedical area taken during the first year and "1" in case of successfully completing all the coursework. For 1200 students, a forecast for passing the exam was received, the quality of which was assessed using the ROC curve and equals to 0.7736 and thus considered adequate.

Researchers note (Troussas et al., 2013), that clustering users into groups with common interests is very useful when learning multiple languages. They used the k-means algorithm because of its simplicity. Authors (Bayazit et al., 2022) use the same clustering algorithm to identify students with low engagement. A known drawback of the algorithm is that the number of clusters is set a priori and does not sufficiently reflect students with a satisfactory level of interaction. Two clustering quality indexes are tested and compared in (Petrovic, 2006). Experimental results comparing the effectiveness of a multiple classifier with the two indexes implemented show that the system using the Silhouette index produces slightly more accurate results than the system that uses the Davies-Bouldin index.

Based on the literature review, it can be concluded that there is significant interest in predicting students' academic performance using machine learning methods. It has been established that academic performance prediction is carried out either through binary classification – whether a student will pass the exam or not, or through regression to predict the potential score for the exam. It has been identified that methods grouping students based on similarities in learning trajectories are not commonly applied.

Quick overview of the initial dataset

The quality of the data of the e-learning platform has a direct impact on the accuracy of predictive models (Qiu et al., 2022). The initial dataset with intermediate and final scores of students for the first semester in the discipline Data Analysis contains 20 features: two text features with the surnames and first names of students; number of a group; eight digital features with students' scores for homework, self-study, term papers and tests posted in the e-learning platform; and the remaining numerical features contain the scores given by the teacher for programming skills, activity in the classroom, as well as final scores.

To visualize numerical features (A. Yu. Vladova et al., 2021) we apply the WordCloud library of the Python language, and built the tag cloud shown in Figure 1. Based on Figure 1, it is evident that the educational institution employs two grading scales: midterm performance is evaluated on a scale from 0 to 5.0, while the final scores are converted to a scale from 0 to 100.0 with maximum scores of 5 and 100 points respectively. The majority of students demonstrate good and excellent knowledge

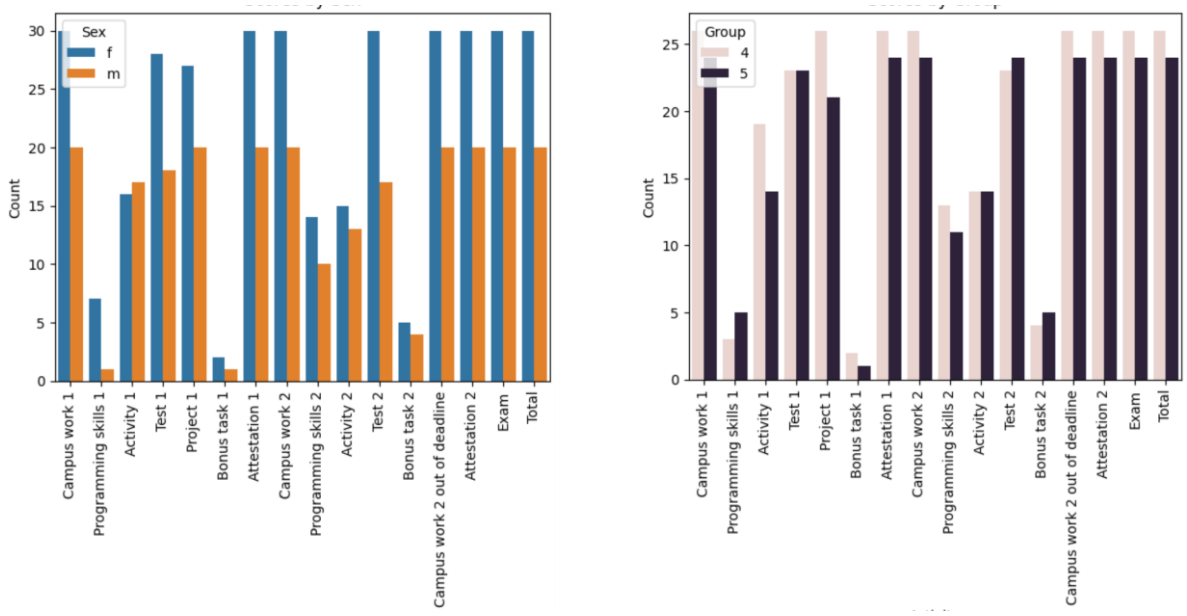

Figure 2. The map of the number of scores for different activities: a) taking into account the students' gender; b) taking into account the students' group

throughout the semester. However, there is a wider variation in the final scores. Therefore, it is necessary to identify the weak points in students' preparation that impact their final performance.

Purpose and objectives of the study

The aim of the study is to enhance students' preparedness by developing models for assessing and predicting students' academic performance based on their current scores. The objectives of this research are:

1) Conduct a statistical analysis of student scores for a particular discipline: assess data imbalance, identify gaps, construct score distribution densities, and explore the correlation matrix of scores.
2) Increase the dimensionality of the feature space by normalizing scores to a common scale and creating new features such as student indexes, ranks, and the differences between scores at different time points, obtained by each student.
3) Predict students at academic risk and evaluate the statistical significance of the features included in the model.
4) Customize consultations for student groups with similar learning trajectories.
5) Forecast semester final scores for each student. This approach involves a comprehensive analysis, modeling, and customization of consultations to effectively improve students' academic performance levels in universities.

Statistical analysis of the initial dataset

It is essential to address the data imbalance highlighted in the primary statistical analysis:

1) Gender imbalance, with women constituting 30% more than men
2) Ibalance in the number and level of scores received by students for various activities. The histograms depicting the number of students assessed for each activity, segmented by gender and group, are presented in Figure 2.

The analysis revealed that male students are more involved in coursework and significantly more active in classes, while a significantly larger number of female students completed the second test. Bonus assignments are challenging for both male and female students. There are missing values in the data shown in Figure 3 because not all students completed the full amount of work. These

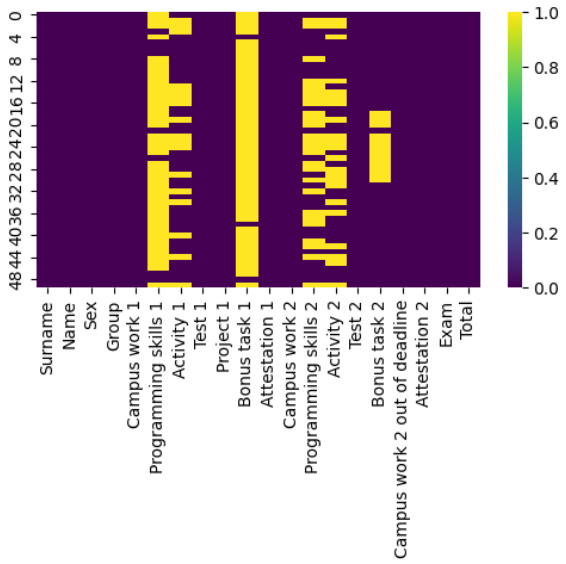

Figure 3. Map of data omissions

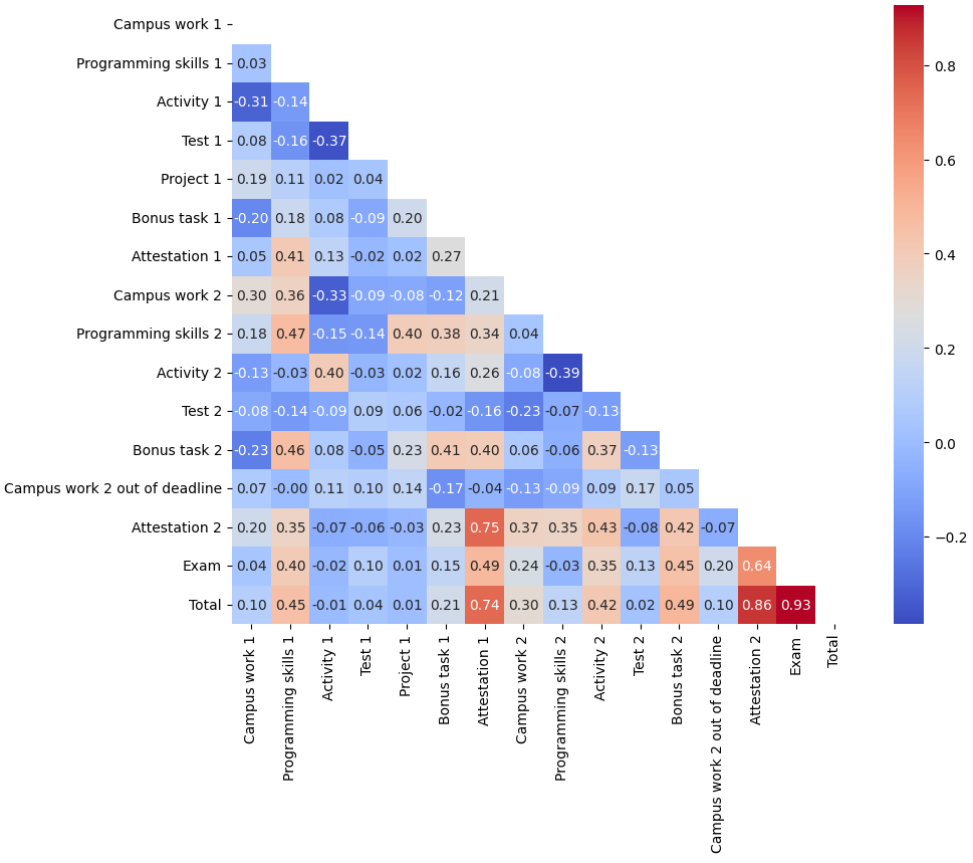

Figure 4. Correlation matrix

missing values are logically replaced with zero scores. From Figure 3, it is evident that there is a positive trend in programming skills but a negative trend in class activity.

Understanding the relationships between features allows for better preparation for the clustering process, eliminating redundant or highly correlated features (Hafsa et al., 2023). The lower triangular correlation matrix in Figure 4 shows that the highest coefficients of linear correlation are

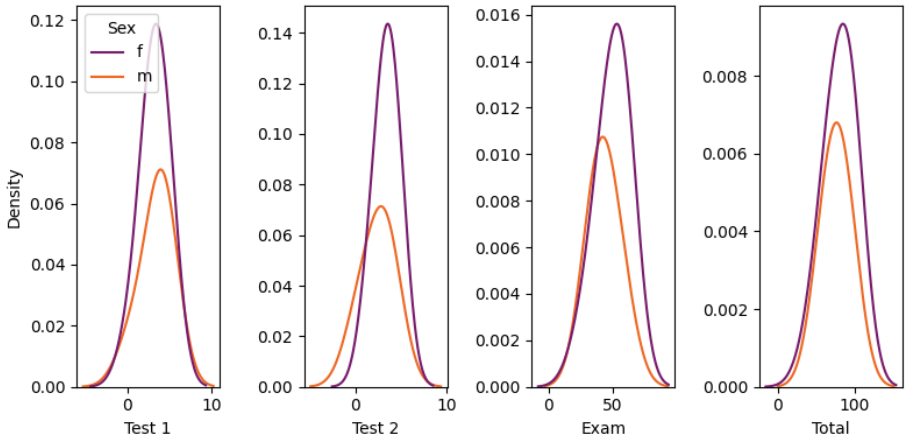

Figure 5. Score distribution densities

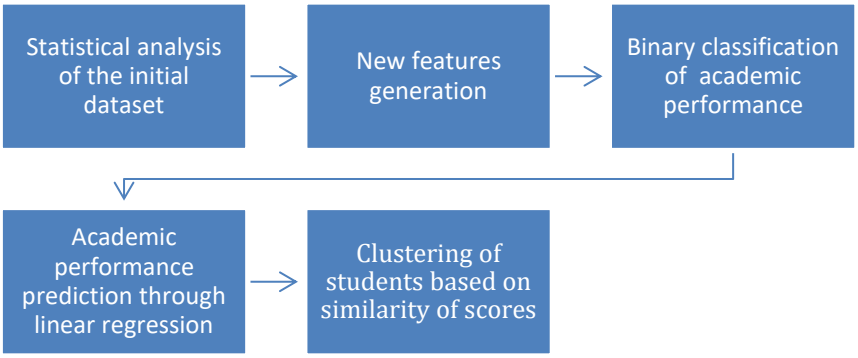

Figure 6. Stages of the methodology

observed between the second module certification and the final score (0.86), as well as between exam scores and the final score (0.93).

We made the density of score distributions across the main control points (tests, exam, final score) close to normal shape shown in Figure 5 using the Kernel density estimation method (Humbert et al., 2022; Węglarczyk, 2018). Moreover, the grading scales for tests vary from 0 to 10, for the exam from 0 to 60, for the final score from 0 to 100. The average score for men for the second test, exam and final score is slightly shifted to the left relative to the average score for women.

METHODS

The proposed methodology for predicting the final score includes five stages shown in Figure 6. The first stage involves performing a statistical analysis to identify imbalances, data omissions, high and low correlations. In the second stage, new features are formed, and the dataset is aggregated by student index and group number, forming a performance for each student. In the third stage, with a sufficient number of scores in the aggregate, a binary classification predictive model for students who passed and did not pass the exam is built along with an error assessment. There are several methods that implement a binary classification and the most effective methods are those derived from linear discriminant analysis (such as Quadratic Discriminant Analysis (QDA), Regularized Discriminant Analysis (RDA) and Logistic regression (Araveeporn, 2023). The choice of logistic regression is dictated by the fact that it does not assume normal distribution of independent variables and homogeneity of variation-covariance matrices. The quality of separation is evaluated using the F1 metric - the harmonic mean of accuracy and completeness:

$$F1 = \frac{TP}{TP+\frac{FP+FN}{2}}, \quad (1)$$

where TP, FP, FN – are true positive, false positive and false negative forecasts.

In the fourth stage performance is forecasted using linear regression. As a result, a trend and a predicted performance score are determined for each student. The quality of prediction is evaluated through mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) (Aissaoui et al., 2020):

$$MAE = \frac{\sum_i^n |y_i - x_i|}{n}, MSE = \frac{1}{n}\sum_i^n (y_i - x_i)^2 \ , RMSE = \sqrt{MSE} \quad (2)$$

where y_i is the prediction and x_i is the true value, n – the observations number.

Finally, the fifth stage entails clustering students based on the similarity of their score set. The quality of clustering is determined by the silhouette coefficient. It measures how well an object matches its cluster compared to other clusters: a value close to 1 indicates that the object is well clustered, a value close to 0 indicates that the object is on the border between two clusters, a negative value indicates incorrect clustering.

The Silhouette Score can be calculated for each feature in the cluster and then averaged for an overall assessment of the clustering quality (Bonaccorso, 2018). The formula for calculating the silhouette coefficient for the individual object i is as follows:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (3)$$

where $a(i)$ — The average distance from feature i to all other features in the same cluster. This value indicates how close the features within the cluster are to each other; $b(i)$ — The minimum average distance from feature i to features in the nearest other cluster. This value indicates how close the feature is to other clusters.

The average value of the silhouette coefficients of all objects is calculated according to the formula:

$$S = \frac{1}{n}\sum_{i=1}^n s(i), \quad (4)$$

where n is the total number of objects.

The proposed method differs from known analogs in that, firstly, the dimensionality of the feature space increases due to the normalization of scores onto a single scale and the creation of new features: the index and rank of students, as well as the changes in performance across various activities for each student. Secondly, students at academic risk are forecasted, and the statistical significance of the features included in the model is evaluated. Thirdly, for each student, the final score for the semester is forecasted using an linear regressive model of academic performance. Fourthly, groups of students with similar learning trajectories are identified for customization of consultations.

The practical significance of this method lies in the possibility of obtaining new knowledge about the learning process.

New features generation

Creating new features improves models in the following aspects: reducing calculation speed or required data volume, enhancing model interpretability, and increasing predictive accuracy (A. Vladova & Shek, 2021) . Based on the names and endings of Russian surnames (Zahoranský & Polasek, 2015), the binary attribute Sex has been added. To anonymize the data (Alier et al., 2021), a feature called Index is introduced, comprised of the first letters of the student's last name, first name, gender, and group number. It was found that test, project, and homework scores range from 0 to 5 points, midterm assessment scores vary from 0 to 22 points, exam scores range from 0 to 60 points, and final scores range from 0 to 100 points. Therefore, all scores are normalized to a range from 0 to 1 by dividing by their respective maximum values. This transformation results in new normalized features that comprehensively characterize an academic performance on a scale from 0 to 1. The distributions of these features are shown in Figure 7.

The next feature contains the student's rank relative to other students in the group and is obtained by calculating the sum of the product of the scores by the weighting coefficients and sorting the results from maximum to minimum shown in Table 1.

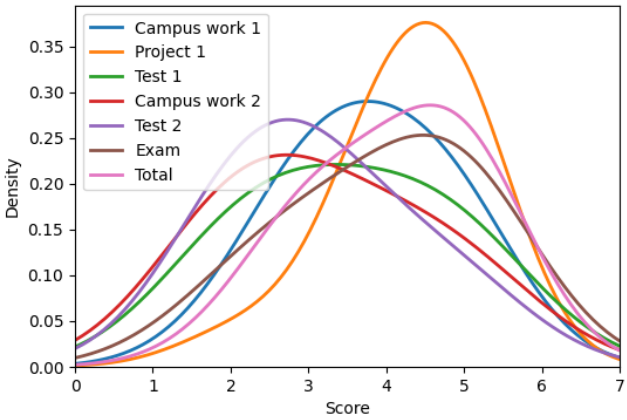

Figure 7. Distribution of normalized scores by main types of academic work

Table 1
Student's rank. Fragment

Index	Rank
BAf5	1
MAf5	2
	.....
HAm4	49
ГAf5	50

Table 2
Academic performance. Fragment

Index	Campus work progress	Programming skills progress	Activity progress	Test progress	Attestation progress
ВДmf5	0.11	0.6	0.2	-0.1	0.13
ВКf5	0.23	0.0	0.0	0.9	0.41
ГAf5	-0.06	0.0	0.0	0.1	0.01
MAf5	0.12	0.0	0.4	0.0	-0.11
	....		....		

To provide a more comprehensive view of a student's academic progress we consider changes in academic performance across various activities shown in Table 2. Negative values indicate a decrease in score in the second part of the semester. This approach can be particularly useful for educators and academic institutions to gain a deeper understanding of student development.

By calculating the differences between normalized scores at different time points (e.g., Campus work 2 - Campus work 1, Attestation 2 – Attestation 1, etc.), these new features effectively capture the change in an academic performance or achievement in specific activities over a period of time, such as a semester.

The article (Shahiri et al., 2015) gives the top four methods for predicting academic performance. The neural network has the highest prediction accuracy (98%), followed by the decision tree (91%). Further, the support vector machine and the K-nearest neighbor machine gave the same accuracy, which is (83%). Finally, the method that has lower prediction accuracy is the naive Bayes method (76%). Since the number of scores for each student in the existing dataset is small, it is inefficient to use a neural network. For the available data, it is rational to use one of binary classification methods.

Binary classification of academic performance

At the first stage we need to classify students into those who will pass or fail the exam. The logistic regression is one of binary classification method applicable when the dependent variable is dichotomous. Let's assume that passing the exam is the target event (A. Yu. Vladova, 2024). There is

a

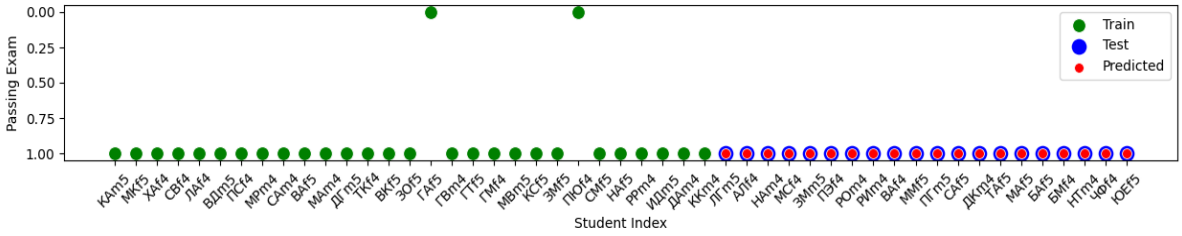

Figure 8. Results of classifying students into those who passed and those who did not pass the exam

Table 3
Feature Importance

Number	Feature	Importance
0	Sex	0.502
5	Test progress	0.475
3	Programming skills progress	0.254
2	Campus work progress	0.189
1	Group	0.169
4	Activity progress	0.125
7	Attestation progress	0.054
6	Bonus task progress	0.009

labeled dataset - students' scores on various tasks (progress in Campus works, Tests, and Attestations) and their score for the exam. The training (80%) and test (20%) datasets are separated from it. The logistic regression model is then trained and evaluated for accuracy as follows: Let X be the vector of input features (students' scores on various tasks), and Y be the binary output (pass/fail the exam). Let's assume that a student has the probability of passing the exam is:

$$P\{y = 1|x\} = f(z), \quad (5)$$

where $z = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n$ are column vectors of the values of the input normalized features x and regression coefficients θ; f(z) – logistic function defined as $f(z) = \frac{1}{1+e^{-z}}$.

Visualization of classification results uses student indexes, where learning outcomes and predicted test results are indicated in different colors shown in Figure 8. The output is a trained logistic regression model capable of predicting whether a student will pass an exam based on their input scores. After training the logistic regression model, the importance of features was estimated by the absolute values of their coefficients shown in Table 3.

Features with higher coefficients are more important in predicting the target variable (Bruce & Bruce, 2017). Therefore, the latter feature can be excluded from model training. In addition, we checked its statistical significance with the Student's test (*How to Do a T-Test in Python | Built In*, n.d.). There is a negative trend, but the effect of project marks on the exam result does not demonstrate statistical significance (t-statistic = -1.99, p > 0.05). In this case, there is no reason to conclude that project scores have a significant impact on exam passing.

Academic performance prediction

The normalized set of input and output features is again broken down into training and test parts. An instance of the linear regression model learns from the training part and makes predictions for the test part. The data is visualized using a scatter plot. Figure 9 showing the exam result for several student's index. To evaluate the performance of the linear regression model, error metrics are calculated on test data: MAE = 0.079, MSE = 0.078, RMSE = 0.088, R squared = 0.89.

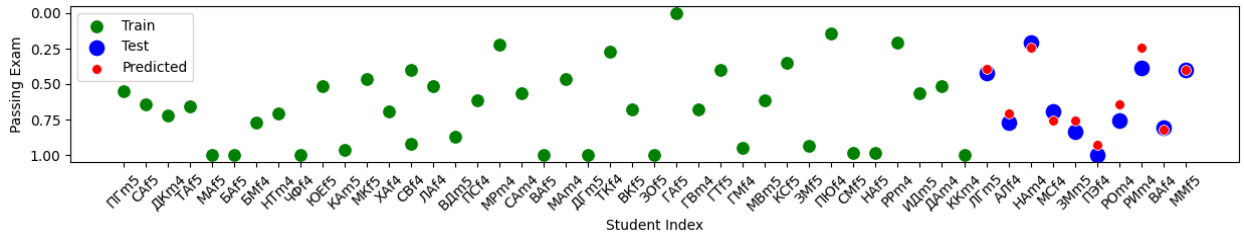

Figure 9. Academic performance prediction

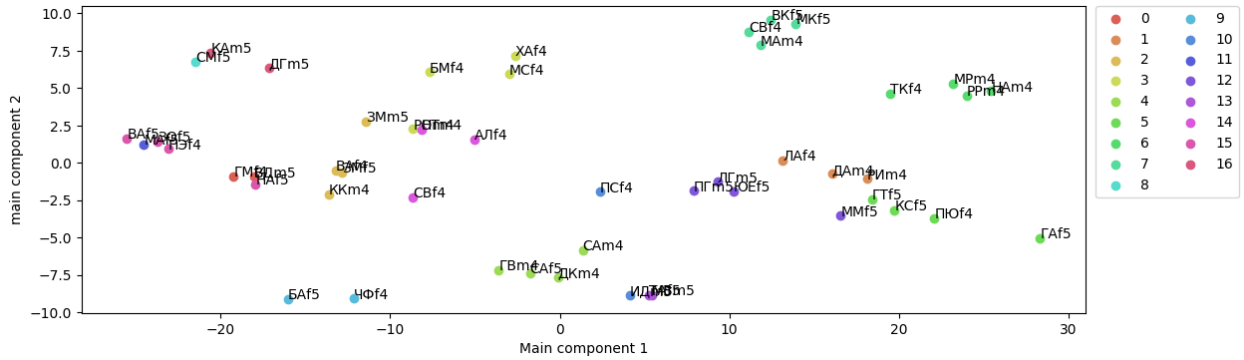

Figure 10. Results of classifying students by similarity of scores

For the available dataset, the model shows high accuracy because the known test and predicted values are close enough, and the errors do not exceed 9%. The t-statistic of -2.83 and a p-value of 0.01 suggest that there may be a significant difference in the Test progress feature between the groups of males and females being compared (Olatunde-Aiyedun, 2021). The low p-value indicates evidence to reject the null hypothesis in favor of a significant difference.

Clustering students by similarity of scores

When creating personalized learning plans, identifying successful/unsuccessful learning strategies, it is useful to identify groups of students with similar sets of assessments (Reiser & Joseph's College, 2017). For this purpose, students are clustered using the k-means method (Wati et al., 2021) on the same labeled dataset. Let X be the vector of input features (students' scores for various tasks) and Y be the output feature (cluster number). The initialization of the mass centers of the clusters is random. The algorithm seeks to minimize the total standard deviation of the cluster points from the centers of these clusters:

$$V = \sum_{i=1}^{k} \sum_{x \in S_i} (x - \mu_i)^2, \quad (3)$$

where k is the number of clusters, S_i are resulting clusters, i=1,2,...., k, and μ_i are centers of mass of all X vectors from the cluster S_i.

The steps are repeated until convergence, that is, until the centroids stop changing significantly or until the maximum number of iterations is reached (Boehmke & Greenwell, 2020). The results of clustering are presented in a two-dimensional plot in Figure 10 using the principal component analysis (Ahmad et al., 2019), which reduces the dimensionality of the data space by converting a large set of features into a smaller one with minimal loss.

To find the optimal number of clusters, three characteristics are calculated: inertia, silhouette index, and Davis-Baldwin index. Inertia is computed as the sum of the squared distances from each data point to its nearest cluster centroid. It shows how grouped the points are within all the clusters in Figure 11. The lower the inertia, the better the model, because more compact and dense clusters usually imply a clearer structure in the data (Rykov et al., 2024). The silhouette index is computed for each data item in a cluster by measuring how close it is to the rest of its cluster compared to the

elements of other clusters. The closer the silhouette index value is to one, the better the clusters are

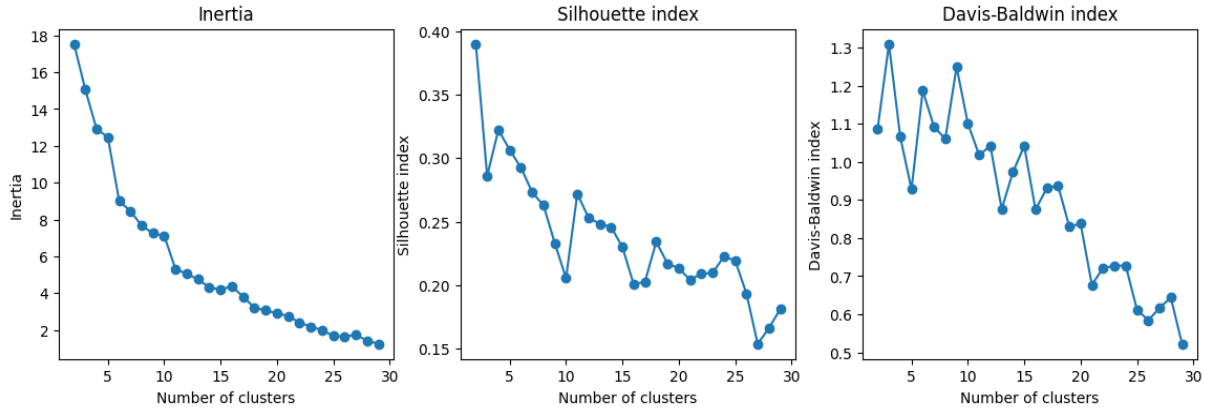

Figure 11. Selection of the optimal number of clusters

separated. The Davis-Baldwin index is computed as the average of the paired distance ratios between the centroids of the clusters and their average intra-cluster distance. The smaller the Davis-Baldwin index, the better the clustering. As a result, 17 clusters of students with similar scores are created with the following characteristics: inertia 890.47, silhouette index 0.28, Davis-Baldwin index 0.83. Analysis of inertia graphs, silhouette and Davis-Baldwin indices showed that the optimal number of clusters varies from 16 to 20.

FINDINGS

To identify methods and key factors influencing academic performance we performed the literature review. To analyze, customize and predict academic performance based on the data from e-learning platforms we offered the multistage methodology. At the first and second stages it applies statistical methods to form new features and improve the predictive ability of models. Thus, correlation analysis revealed a strong relationship between a number of features. Therefore, the dynamic features introduced into the feature space, taking into account the change in academic performance over time. The problems of predicting exam grades, classifying students into passing and non-passing an exam, as well as clustering students by sets of grades are solved at the third, forth and fifth stages consequentially. The results of the classification of exam grades are as follows: the estimate of the harmonic mean value of accuracy and completeness for the initial data F1 is 82 %. The linear regression model demonstrated the following error values: MAE = 0.1, MSE = 0.02, RMSE = 0.1, R2 = 0.7.

DISCUSSION

Classifying, clustering, and predicting academic performance can be useful for multiple stakeholders such as teachers, students, and institutions. For teachers, these tools help identify at-risk students, adapt curricula, and design targeted interventions. Students benefit by gaining insights into their performance trends, enabling better planning of study strategies and group work. Institutions can use these models to identify program-wide trends, evaluate curriculum effectiveness, and allocate resources to address systemic issues.

This article explores the use of machine learning methods to predict student performance, emphasizing several critical steps in the process. To ensure that all features are on a comparable scale, missing data is addressed, and normalization is applied. Key academic components—such as homework, projects, midterm scores, and test scores—serve as predictors, while new features are created to enhance the models' predictive power. The study employs clustering to analyze student behavior, multiple linear regression for performance prediction, and logistic regression for binary classification, such as pass/fail outcomes. The models are evaluated using metrics like mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) to assess their accuracy. Additionally, Principal Component Analysis (PCA) and Kernel Density Estimation (KDE) are used to visualize the data, providing deeper insights into its structure and distribution. Together, these methods offer a robust framework for understanding and predicting academic performance.

At the same time, the proposed multi-method has a number of limitations. For example, when performing the first two stages, which include statistical analysis and generation of additional features, it is necessary to make some efforts to normalize the data, set a threshold for excluding strongly correlated features, and correctly specify pairs of different-temporal features of the same name that are converted into dynamic ones. In addition, when using methodology in universities of a humanitarian orientation, the Programming skills attribute can be replaced, for example, by attendance.

At the last stage, the k-means clustering method is used to divide students into groups. This is a simple and straightforward method, which, unlike more modern clustering methods (e.g., DBSCAN (Vladova, 2024), DBCLASD (Sheikholeslami & Zhang, 1998), WaveClaster (Sheikholeslami & Zhang, 1998)) involves a separate expert study on the number of clusters. This study includes an estimate of the inertia, silhouette index, and Davis-Baldwin index and results in a recommended cluster count interval. Within this interval, the educational manager must select one number – the exact number of clusters. Such a choice requires a certain expertise from the decision-maker, but at the same time allows him to take into account the administrative restrictions on the number of groups of students studying in different programs.

In the subsequent study, it is proposed to exclude from further consideration students at academic risk identified at the first stage (Shou et al., 2024), and also to investigate the impact on academic performance of signs of IP address coincidences when doing homework (Komosny & Rehman, 2022), duration of work, start and end time of work. In addition, it is necessary to develop dashboards that greatly facilitate model settings and decision-making for managers and teachers of educational institutions.

CONCLUSION

The literature review highlights various approaches to predicting student academic performance using machine learning and statistical methods. Researchers emphasize the importance of identifying key factors influencing performance, such as prior academic scores, and engagement in e-learning platforms. Various models, including regression analysis and classification techniques, have been utilized, demonstrating mixed success rates, while suggesting the need for further feature selection and data transformation to enhance predictive accuracy. The weaknesses of the approaches include insufficient accuracy of the models, the use of qualitative features, and the influence of experts

The study carried out a comprehensive statistical analysis of students' scores for a math discipline. This involved assessing data imbalance, identifying gaps, constructing score distribution densities, and exploring the correlation matrix of scores. These analyses provide valuable insights into the distribution and relationships of student scores, laying a strong foundation for further modeling and predictions.

The study changed the dimensionality of the feature space by normalizing scores to a common scale and creating new features such as student indexes, ranks, and differences between scores at different time points. This expansion of the feature space enhances the richness of the dataset and can potentially lead to more robust and accurate predictive models.

By developing models to predict students at academic risk and evaluating the statistical significance of the included features, the study addresses the crucial issue of identifying and supporting students who may be at risk of underperforming. This proactive approach can help institutions tailor interventions and support to students who need it most.

The study's plan to customize consultations for student groups with similar learning trajectories reflects a student-centric approach to enhancing academic performance. By recognizing the diverse needs of student cohorts and tailoring support accordingly, the study aims to foster a more personalized and effective learning environment.

The approach of predicting exam scores for individual students demonstrates a commitment to providing comprehensive support beyond mere assessment. By leveraging analysis, modeling, and customization of consultations, the study aims to proactively improve students' academic performance levels within university settings.

The proposed multi-method to the analysis of data from electronic platforms shows a picture of student engagement close to reality. The progress track, predictive assessment and clustering allows educational managers and teachers to assign consultations to groups of students at academic risk and with deteriorating academic performance.

AUTHOR'S DECLARATION

Authors' contributions	All authors contributed to the concept and design of the study as well as result of its. The AYuV: initiator of the main idea and concept of the study, data analysis and validation, KMB: Review and Editing.
Funding Statement	This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors
Availability of data and materials	All data are available from https://github.com/avladova/Student-performance-prediction .
Competing interests	The authors declare that the publishing of this paper does not involve any conflicts of interest. This work has never been published or offered for publication elsewhere, and it is completely original.

BIBLIOGRAPHY

Ahmad, N. B., Alias, U. F., Mohamad, N., & Yusof, N. (2019). Principal Component Analysis and Self-Organizing Map Clustering for Student Browsing Behaviour Analysis. *Procedia Computer Science*, *163*, 550–559. https://doi.org/10.1016/J.PROCS.2019.12.137

Aissaoui, O., Madani, Y., Oughdir, L., Dakkak, A., & EL ALLIOUI, Y. (2020). *A Multiple Linear Regression-Based Approach to Predict Student Performance* (pp. 9–23). https://doi.org/10.1007/978-3-030-36653-7_2

Alier, M., Casañ Guerrero, M. J., Amo, D., Severance, C., & Fonseca, D. (2021). Privacy and e-learning: A pending task. *Sustainability (Switzerland)*, *13*(16). https://doi.org/10.3390/SU13169206

Araveeporn, A. (2023). Comparison of Logistic Regression and Discriminant Analysis for Classification of Multicollinearity Data. *WSEAS TRANSACTIONS ON MATHEMATICS*, *22*, 120–131. https://doi.org/10.37394/23206.2023.22.15

Arzamastsev, S. A., Bgatov, M. V., Kartysheva, E. N., Derkunskii, V. A., & Semenchikov, D. N. (2018). Forecasting Subscriber Churn: Comparison of Machine Learning Methods. *Computer Tools in Education*, *5*, 5–23. http://cte.eltech.ru/ojs/index.php/kio/article/view/1542

Bayazit, A., Ilgaz, H., Gönüllü, İ., & Erden, Ş. (2022). Profiling students via clustering in a flipped clinical skills course using learning analytics. *Medical Teacher*, *45*(7), 724–731. https://doi.org/10.1080/0142159x.2022.2152663

Boehmke, B., & Greenwell, B. (2020). Hands-on Machine Learning with R. In *CRC Press*. https://www.routledge.com/Hands-On-Machine-Learning-with-R/Boehmke-Greenwell/p/book/9781138495685

Bonaccorso, Giuseppe. (2018). Machine Learning Algorithms. In *Packt Publishing: Vol. 2nd ed*. Packt Publishing Ltd. https://www.oreilly.com/library/view/machine-learning-algorithms/9781789347999/

Bruce, P., & Bruce, A. (2017). *Practical Statistics for Data Scientists*. O'Reilly. https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/ch04.html

Elisabeta, P. M., & Alexandru, M. R. (2018). Comparative Analysis of E-Learning Platforms on The Market. *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 1–4. https://doi.org/10.1109/ECAI.2018.8679004

Hafsa, M., Wattebled, P., Jacques, J., & Jourdan, L. (2023). E-learning recommender system dataset. *Data in Brief*, *47*, 108942. https://doi.org/https://doi.org/10.1016/j.dib.2023.108942

How to Do a T-Test in Python | Built In. (n.d.). Retrieved March 8, 2024, from https://builtin.com/data-science/t-test-python

Humbert, P., Bars, B. Le, & Minvielle, L. (2022). Robust Kernel Density Estimation with Median-of-Means principle. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th International Conference on Machine Learning* (Vol. 162, pp. 9444–9465). PMLR. https://proceedings.mlr.press/v162/humbert22a.html

Ibrahim, Z., & Rusli, D. (2007). Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression. *21st Annual SAS Malaysia Forum*.

Kahramanoğlu, R. (2018). Analysis of Changes in the Affective Characteristics and Communicational Skills of Prospective Teachers: Longitudinal Study. *International Journal of Progressive Education*, *14*(6), 177–199. https://doi.org/10.29329/IJPE.2018.179.14

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC Dropout over Weeks Using Machine Learning Methods. In C. Rose & G. Siemens (Eds.), *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs* (pp. 60–65). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-4111

Komosny, D., & Rehman, S. U. (2022). A Method for Cheating Indication in Unproctored On-Line Exams. *Sensors (Basel, Switzerland)*, *22*(2). https://doi.org/10.3390/S22020654

Liu, M., & Yu, D. (2023). Towards intelligent E-learning systems. *Education and Information Technologies*, *28*(7), 7845–7876. https://doi.org/10.1007/s10639-022-11479-6

Olatunde-Aiyedun, T. (2021). Student Teachers' Attitude towards Teaching Practice. *International Journal of Culture and Modernity*, *8*, 6–17. http://ijcm.academicjournal.io/index.php/ijcm/article/download/59/58

Oluwadele, D., Singh, Y., & Adeliyi, T. (2023). E-Learning Performance Evaluation in Medical Education—A Bibliometric and Visualization Analysis. *Healthcare*, *11*, 232. https://doi.org/10.3390/healthcare11020232

Petrovic, S. V. (2006). *A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b2db00f73fc6b97ebe12e97cfdaefbb2fefc253b

Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., Jiang, B., & Chen, P. (2022). Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*, *12*(1), 453. https://doi.org/10.1038/s41598-021-03867-8

Reiser, E., & Joseph's College, S. (2017). Blending Individual and Group Assessment: A Model for Measuring Student Performance. *Journal of the Scholarship of Teaching and Learning*, *17*(4), 83–94. https://doi.org/10.14434/JOSOTL.V17I4.21938

Rykov, A., De Amorim, R. C., Makarenkov, V., & Mirkin, B. (2024). Inertia-Based Indices to Determine the Number of Clusters in K-Means: An Experimental Evaluation. *IEEE Access*, *12*, 11761–11773. https://doi.org/10.1109/ACCESS.2024.3350791

Shahiri, A., Husain, W., & Abdul Rashid, N. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Sheikholeslami, G., & Zhang, A. (1998). A Multi-Resolution Clustering Approach for Very Large Spatial Databases *. *Proceedings of the 24th VLDB Conference*. https://www.vldb.org/conf/1998/p428.pdf

Shou, Z., Xie, M., Mo, J., & Zhang, H. (2024). Predicting Student Performance in Online Learning: A Multidimensional Time-Series Data Analysis Approach. *Applied Sciences*, *14*(6). https://doi.org/10.3390/app14062522

Troussas, C., Virvou, M., & Alepis, E. (2013). Comulang: towards a collaborative e-learning system that supports student group modeling. *SpringerPlus*, *2*(1), 387. https://doi.org/10.1186/2193-1801-2-387

Urrutia-Aguilar, M., Fuentes-Garcia, R., Martinez, D., Beck, E., Ortiz, S., & Guevara-Guzmán, R. (2016). Logistic Regression Model for the Academic Performance of First-Year Medical Students in the Biomedical Area. *Creative Education*, *07*, 2202–2211. https://doi.org/10.4236/ce.2016.715217

Vladova, A. Yu. (2024). Developing group and individual performance paths based on e-learning platform data. *Large-Scale Systems Control (UBS)*, *111*, 179–196. https://doi.org/10.25728/ubs.2024.111.7

Vladova, A., & Shek, E. (2021). Data preprocessing for machine analysis of sales representatives' key performance indicators. *Business Informatics*, *15*(3), 48–59. https://doi.org/10.17323/2587-814X.2021.3.48.59

Vladova, A. Yu., Vladov, Yu. R., & Yakimov, A. I. (2021). Visualizing Results of Promoting Campaigns. *2021 14th International Conference Management of Large-Scale System Development (MLSD)*, 1–4. https://doi.org/10.1109/MLSD52249.2021.9600205

Wati, M., Rahmah, W. H., Novirasari, N., Haviluddin, Budiman, E., & Islamiyah. (2021). Analysis K-Means Clustering to Predicting Student Graduation. *Journal of Physics: Conference Series*, *1844*(1), 012028. https://doi.org/10.1088/1742-6596/1844/1/012028

Węglarczyk, S. (2018). Kernel density estimation and its application. *ITM Web of Conferences*, *23*, 00037. https://doi.org/10.1051/ITMCONF/20182300037

Yadav, N., & Deshmukh, S. (2023). *Prediction of Student Performance Using Machine Learning Techniques: A Review* (pp. 735–741). https://doi.org/10.2991/978-94-6463-136-4_63

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Yang, S. J. H., Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Ogata, H., & Lin, A. J. Q. (2018). Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis. *Journal of Information Processing*, *26*, 170–176. https://doi.org/10.2197/IPSJJIP.26.170

Ye, M., Sheng, X., Lu, Y., Zhang, G., Chen, H., Jiang, B., Zou, S., & Dai, L. (2022). SA-FEM: Combined Feature Selection and Feature Fusion for Students' Performance Prediction. *Sensors*, *22*(22), 8838. https://doi.org/10.3390/s22228838

Zafar, B., Alhassan, A., & Mueen, A. (2020). Predict Students' Academic Performance based on their Assessment Grades and Online Activity Data. *International Journal of Advanced Computer Science and Applications*, *11*. https://doi.org/10.14569/IJACSA.2020.0110425

Zahoranský, D., & Polasek, I. (2015). Text search of surnames in some Slavic and other morphologically rich languages using rule based phonetic algorithms. *IEEE Transactions on Audio, Speech and Language Processing*, *23*(3), 553–563. https://doi.org/10.1109/TASLP.2015.2393393

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.698490