UDC 004.8
DOI: 10.15587/1729-4061.2024.313333

DETERMINING THE INFLUENCE OF DATA ON WORKING WITH VIDEO MATERIALS ON THE ACCURACY OF STUDENT SUCCESS PREDICTION MODELS

Vladyslav Pylypenko
Corresponding author
PhD Student
Department of Computer Science**
E-mail: software.proger@gmail.com
Volodymyr Statsenko
Doctor of Technical Sciences, Associate Professor*
Tetiana Bila
PhD, Associate Professor*
Dmytro Statsenko
PhD, Associate Professor*
*Department of Computer Engineering
and Electromechanics**
**Kyiv National University of Technologies and Design
Mala Shyianovska (Nemyrovycha-Danchenka) str., 2,
Kyiv, Ukraine, 01011

The object of this study is models for predicting students' success, constructed on the basis of machine learning methods. The paper reports results of research into the problem of improving their accuracy by expanding the data set for training the specified models. The most available are data on student actions, which are automatically collected by learning management systems. Entering additional information about students' work increases time and resources but allows the improvement of the accuracy of the models. In the study, information about students' work with video materials, particularly the number and duration of views, was entered into the original data set. To automate the collection of this data, the plugin for the Moodle system has been developed, which stores information about user's actions with the video player and the duration of watching video materials in the database. Model training was carried out using Naive Bayes (NB), logistic regression (LR), random forest (RF), and neural networks (NN) algorithms with and without video data. For the models using video viewing data, accuracy increased by 10 %, balanced accuracy by 15 %, and overall performance, expressed as area under the curve (AUC), increased by 14 %. The highest prediction accuracy, with a difference of 1.8 %, was obtained by models built using RF algorithms – 87.1 % and NN – 85.3 %. At the same time, the accuracy of the models obtained by the NB and LR algorithms was 70.7 % and 76.5 %. The increase in accuracy for them was 2.3 % and 8.1 %, respectively. Analysis of calculations confirms the assumption that students' work with educational video materials is correlated with their success. The results make it possible to find a reasonable compromise between model development costs and its accuracy at the stage of data preparation for model training

Keywords: success prediction, random forest, logistic regression, neural networks, naive Bayes

Received date 05.08.2024
Accepted date 09.10.2024
Published date 30.10.2024

How to Cite: *Pylypenko, V., Statsenko, V., Bila, T., Statsenko, D. (2024). Determining the influence of data on working with video materials on the accuracy of student success prediction models. Eastern-European Journal of Enterprise Technologies, 5 (4 (131)), 52–62. https://doi.org/10.15587/1729-4061.2024.313333*

1. Introduction

Education is a key factor that determines the future opportunities and career path of students.

Along with the spread of computer technologies, e-learning is gradually becoming an alternative to traditional forms of education [1]. This transition from classic classroom methodology to online platforms and digital resources not only enables flexibility and accessibility of education but also allows for the integration of innovative teaching methods, such as interactive courses, audio and video lectures, as well as virtual laboratories [2, 3]. Owing to learning management systems (LMS), numerous advantages have been obtained: the ability to access educational materials from anywhere in the world, adaptation of the educational process to the individual needs of students, and the use of state-of-the-art technologies for monitoring and evaluation [4].

One such system is Moodle; it has a high level of recognition in many educational institutions, is free, has a wide range of active courses available in many languages [5]. Given the growing amount of data generated in online learning, the ability to analyze this data and use it to predict academic achievement allows timely identification of students who need additional support. Accurate prediction of success is a complex process as it depends on many factors, in particular attendance, involvement and activity in the educational process, working with educational materials and completing tasks [6].

Machine learning methods are mainly used to predict student performance. The most common classification algorithms are logistic regression (LR), decision tree (DT), naive Bayes classifier (NB), support vector method (SVM), random forest (RF), neural networks (NN) [7]. Achieving high accuracy in predicting success is a complex and multifaceted task due to the large number of factors that influence it. Data quality is one of the main factors and in fact the basis of any successful forecasting. The first thing that arises when constructing a model is the collection of data for its training. The most accessible way of obtaining them under real conditions is the analysis of information collected by LMS, in particular, Moodle. Next, there is the question of increasing the amount of available data by obtaining various additional parameters of students' interactions with educational materials in the Moodle environment. One of the problems is obtaining information about working with video materials since the basic functionality of the Moodle system does not store information about the viewing parameters of such materials. In this context, the development and implementation of a plug-in for obtaining additional data on students' interaction with educational materials and their monitoring at the university becomes particularly important [8].

Scientific research on this topic is important because student success is a strong indicator of the quality of educational services provided by an educational institution. And the results of predicting success allow students to build their plans for achieving goals, and lecturers get the opportunity to identify students at risk and intervene in time [9]. Due to the prediction of student success using machine learning methods, problems with success can be detected at an early stage. This highlights the importance of finding and applying better classification algorithms for more accurate results. Thus, research aimed at improving and improving the accuracy of performance prediction models is certainly of great importance to education.

2. Literature review and problem statement

Intelligent analysis of educational data has become an effective tool for researching hidden relationships in educational data and predicting students' educational achievements. And the use of machine learning made it possible to build effective models to perform rapid forecasting. Nevertheless, it is important to understand that the accuracy of predicting success depends not only on algorithms and machine learning methods but also on the number and weight of parameters included in the model.

Work [10] reports the results of research on predicting student success based on final exam grades. That is, the success factor is a set of 60 or more points by a student. The data was taken from the Student Information System (SIS). In these records, midterm grades, final exam grades, and midterm course grades are collected from 1,854 students. A total of three types of parameters were used: grades for midterm exams, grade data from the department, and grade data from the faculty. The accuracy of random forest (RF), nearest neighbor (NN), support vector (SVM), logistic regression (LR), naive Bayes (NB) and *k*-nearest neighbor (kNN) algorithms was calculated and compared, which was: 74.6 %, 74.6 %, 73.5 %, 71.7 %, 71.3 %, and 69.9 %, respectively. It is shown that the random forest algorithm was able to achieve the highest classification accuracy of 74.6 %. The calculated area under the curve (AUC) value for the RF, NN, SVM, LR, NB, and kNN algorithms was 86 %, 86.3 %, 80.4 %, 82.6 %, 81 %, and 81 %, respectively. Prediction accuracy was assessed using ten-fold cross-validation. Using a large number of algorithms to determine the best one is a good solution. However, issues related to the limitation of the data, from which only grades for intermediate results and exams are available, remain unresolved. Therefore, when predicting success at an early stage, when there are no grades yet, it is unlikely to achieve significant accuracy.

Work [11] gives the results of research on predicting the success of students based on class attendance. As data for training and training the model, data on grades and attendance at laboratory works, lectures and practical classes, which were taken from the LMS Moodle, were used. The accuracy of random forest (RF), support vector (SVM), and logistic regression (LR) algorithms was calculated and compared, which was 80 %, 79 %, and 79 %, respectively. The calculated area under the curve (AUC) value for the RF, SVM, and LR algorithms was 73 %, 66 %, and 70 %, respectively. It is shown that the random forest algorithm was able to achieve the highest classification accuracy of 73 %. Attendance is an indicator that directly affects success but cannot be the only sufficient factor. Therefore, issues related to data limitations remain unresolved, as even with 100 % attendance, a student may have low scores and fail the session. Adding other types of student interaction data may be an option to overcome related difficulties.

This is the approach used in work [12], in which the following were used as data for training and learning the model: grades for lectures, tests and laboratory work, and watched videos. To predict success, only the random forest algorithm was used to build the model. It is shown that the obtained model was able to predict failure with an accuracy of 96.3 %. In the work, they used 3-fold cross-validation, repeated 5 times, and only then built a random forest model with centered and scaled data. The absence of a constructed ROC curve and calculation of the area under the curve (AUC) does not give a clear understanding of the overall efficiency of the obtained model. Unsolved issues related to the parameter are the number of viewed videos, namely how it is calculated. Is there really a check that the video is watched to the end, and not just opened by the student? The model shows high accuracy, but it is not clear what the prediction result would be at an early stage. There is also a lack of calculations of the quality and efficiency of the model and a comparison of the accuracy of its prediction with other machine learning algorithms.

In work [13], success prediction was carried out with the following algorithms: random forest (RF), logistic regression (LR), neural networks (NN), gradient extended decision trees (XGBoost). It is shown that the forecasting accuracy was 90 %, 90 %, 87 %, and 84 %, respectively. A comparison of the evaluation scores shows better performance for the neural network and gradient descent boosting tree algorithms compared to logistic regression and random forest. The research hypothesis is that learning success or failure could be predicted using student activity data from LMS Moodle logs. Gender, number of files downloaded, files viewed, modules completed, number of platform visits, activity per day, and other activity data from activity logs in the LMS were used as data for training the model. The results of the study confirm the hypothesis that academic success can be predicted using machine learning algorithms based on data obtained during students' interaction with e-learning platforms. Unsolved issues are related to the fact that reports on user activity do not have data on the actual interaction of the user with them. And only the very fact of switching to an activity or discovery, for example, watching a lecture or an exercise. An option to overcome related difficulties may be to add new data and test the results of the prediction accuracy on a larger number of different samples.

In [14], a model based on a multi-level Perceptron neural network was used to predict success, which was trained to predict students' success in a blended learning environment. Predicted student success is based on four learning activities: email communication, collaborative content creation using a wiki, content interaction as measured by file views, and self-assessment through online quizzes. It is shown that the model predicted the success of students with a correct classification rate (CCR) with an accuracy of 98.3 %. The obtained accuracy is also confirmed by the constructed ROC curve, the value of the calculated area under the curve (AUC) is 98.9 %. For comprehensive assessment and comparison of forecasting results, it is advisable to add other classification algorithms. It is necessary to check the results of the prediction accuracy on a larger amount of data and other algorithms, since the accuracy in this case may change.

Work [15] reports the results of research on predicting student success using academic data and records from the learning management system, which are correlated with success or failure in the course. Six algorithms (GBT, RF, DT, LR, NB, SVM) were used, with model training at three different stages over a two-year course. The models were tested on the records of 394 students from 3 courses. It is shown that Random Forest gave the best results with 84.47 % F1 score in experiments, followed by Decision Tree that produced similar results. Unlike previous studies, this one took into account data from 3 courses, which contributes to a more accurate prediction of the model of students who are at risk of dropping out. But the issue related to the sufficiency of the used data for accurate prediction of success in the early stages of education remained unresolved.

Paper [16] gives the results of research on the development of accurate prediction of the results of the course of students, whether they will pass it or not at all. Unlike previous studies, this study considered demographic, assessment, and student interaction data to provide comprehensive predictions. Logistic regression and random forest were used to develop forecasting models. The accuracy of the models was assessed based on four-class classification (prediction of four possible outcomes) and two-class classification (prediction of pass or failure). It has been shown that simple metrics such as a student's activity level on a given day can be as effective as more complex combinations of data or personal information in predicting student performance. The logistic regression model achieved an accuracy of 72.1 % for four-class classification and 92.4 % for two-class classification. The random forest classifier achieved an accuracy of 74.6 % for four-class classification and 95.7 % for two-class classification. This predictive approach provides insight into course student outcomes, offering valuable information to improve student engagement in online learning environments.

In the reviewed literature, various machine learning algorithms are used, and models are built that have a high calculated accuracy of predicting success. However, most use only data from LMS Moodle logs and point values (grades for subjects, tests, and modules) obtained from previous academic periods. Expanding the data set for training is associated with additional resource costs for their collection and processing. At the same time, the duration of model training process increases. This allows us to state that it is appropriate to conduct a study of the impact of a data set on students' work with video materials on the accuracy of success prediction models. The results of such research will make it possible to make informed decisions about the selection of data sets for training machine learning models.

3. The aim and objectives of the study

The purpose of our study is to determine the impact of data on working with video materials on the accuracy of models for predicting students' success. This will make it possible to make informed decisions about expanding the dataset for training machine learning models using NB, LR, RF, and NN algorithms.

To achieve this goal, the following tasks were set:

– to prepare a set of initial data from the Moodle system for training models using the NB, LR, RF, and NN algorithms, conduct training, determine the accuracy of the obtained models;

– to expand the set of initial data with information about the work of students with video materials, to train models according to the NB, LR, RF, and NN algorithms, to determine the accuracy of the obtained models;

– to calculate the influence of information about students' work with video materials on the accuracy of models for predicting students' success.

4. The study materials and methods

4. 1. The object and hypothesis of the study

In this work, the object of research is the prediction of student success using machine learning with algorithms of random forest, logistic regression, neural networks, and naive Bayes.

The hypothesis of the study assumes that the use of data on the viewing parameters of video materials could increase the accuracy of machine learning models used to predict student success.

Assumptions and simplifications accepted in the research process – all core learning materials are distributed through a learning management system. The available data from LMS Moodle activity logs about task viewing (opening a file or document) do not contain complete information about the student's actual work with the document. Data about real interaction with educational materials, such as: duration of viewing, status of whether you finished the video, the number of stops and rewinds will give more insight into how the student actually worked. Constructing a plug-in for the Moodle system, which will allow collecting data on students' interaction with educational materials, will make it possible to obtain objective data about students' work with these materials.

4. 2. Output data for model training

This study uses a dataset taken from the Moodle database. Grades, student attendance, as well as data on interaction with educational video materials were exported in csv format. In order to check the increase in the accuracy of predicting success due to the data of interaction with video materials, the data were divided into two sets. Data on interactions with video materials were extracted from the first set, the general view is given in Table 1.

The second set contained all data on grades, attendance, and interaction of students with educational video materials, the general view is given in Table 2.

Table 1

A fragment of the table with data for the prediction of the first set

id	DiscMark	LectVisit	PractVisit	LabVisit	TotalVisit
...	...	...	...	...	...
2316	82	83	–1	83	83
2317	51	25	–1	33	29
...	...	...	...	...	...

Note: id is a unique user identifier;
DiscMark – grade for discipline;
LectVisit – percentage of lecture attendance;
PractVisit – percentage of attendance at practical classes;
LabVisit – the percentage of attendance at laboratory classes;
TotalVisit – total visitation percentage.

Table 2

Fragment of a table with input data for forecasting the second set

id	Disc Mark	Lect Visit	Pract Visit	LabVisit	Total Visit	Duration	Play Count	Pause Count	Stop Count	Complete
...	...	...	...	...	...	...	...	...	...	...
2316	82	83	–1	83	83	27	1	1	0	1
2317	51	25	–1	33	29	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...

Note: id is a unique user identifier;
DiscMark – grade for discipline;
LectVisit – percentage of lecture attendance;
PractVisit – percentage of attendance at practical classes;
LabVisit – the percentage of attendance at laboratory classes;
TotalVisit – total attendance percentage;
Duration – video viewing duration (minutes);
PlayCount – the number of clicks on the Play button;
PauseCount – the number of Pause button clicks;
StopCount – the number of clicks on the Stop button;
Complete – the status of completion of watching the video to the end (1 – yes, 0 – no).

4. 3. Construction of forecasting models

Before performing model training, the data set was divided into training and test samples. In order to test how well a model trained on a training sample can predict classes of new data. The volume of data taken for processing was 2599 records of user samples, which were distributed in the ratio of 520/2079, of which the training sample contained 2079, and the test sample 520. Dividing the data into training and test samples helps avoid overtraining the model [17]. Assessment and quality control of models was carried out on the basis of a test sample. The construction of predictive models was performed in the PyCharm programming environment in the Python programming language. The scikit-learn library [18] was used to build the models. The following libraries were used to construct graphs: seaborn and matplotlib [19]. Processing of tabular data was performed using the pandas library, and numerical data – by using numpy [20].

4. 4. Evaluation of accuracy of models

The following indicators were selected as the main criteria for model performance: overall accuracy, accuracy, balanced accuracy, sensitivity, specificity, *F*1*Score*, area under the curve (AUC), and ROC curve. These indicators are calculated on the basis of the error matrix [21].

Overall accuracy shows what percentage of examples were correctly classified. It refers to the proportion of correct predictions, which include true positives and true negatives. The expression for determining the overall accuracy can be written in the form of the following formula:

$$Accuracy=\left(TP+TN\right)/\left(TP+TN+FP+FN\right), \qquad (1)$$

where *TP* (true positives) is the number of correctly predicted positive classes; *TN* (true negatives) – the number of correctly predicted negative classes; *FP* (false positives) – number of incorrectly predicted positive classes; *FN* (false negatives) is the number of incorrectly predicted negative classes.

Sensitivity makes it possible to determine the ability of the model to detect positive cases. The expression for determining the sensitivity can be written in the form of the following formula:

$$Sensitivity=TP/\left(TP+FN\right), \qquad (2)$$

where *TP* (true positives) – the number of correctly predicted positive classes; *FN* (false negatives) – the number of incorrectly predicted negative classes.

Specificity makes it possible to determine the ability of the model to detect negative cases. The expression for determining specificity can be written in the form of the following formula:

$$Specificity=TN/\left(TN+FP\right), \qquad (3)$$

where *TN* (true negatives) is the number of correctly predicted negative classes; *FP* (false positives) is the number of incorrectly predicted positive classes.

Accuracy measures what proportion of predicted positives are true positives. It shows how accurate the model is in predicting positive classes. The expression for determining accuracy can be written in the form of the following formula:

$$Precision=TP/\left(TP+FP\right), \qquad (4)$$

where *TP* (true positives) is the number of correctly predicted positive classes; *FP* (false positives) is the number of incorrectly predicted positive classes.

*F*1*Score* makes it possible to evaluate the model in cases where it is important to simultaneously minimize false positive and false negative predictions. The expression for definition can be written in the form of the following formula:

$$F1Score=\frac{2*Precision*Sensitivity}{Precision+Sensitivity}. \qquad (5)$$

Balanced accuracy provides an estimate of the overall performance of a binary classifier model, taking into account the balance between data classes. The expression for determining the balanced accuracy can be written in the form of the following formula:

$$BalancedAccuracy=\left(Sensitivity+Specificity\right)/2. \qquad (6)$$

The area under the curve (AUC) parameter was used to evaluate the overall effectiveness of the model regardless of the choice of threshold value [22]. It is calculated as the area under the ROC curve and takes values in the range from 0 to 1. The greater the value to 1, the better the quality of the classification model. The expression for determining the AUC takes the following form:

$$AUC = \sum_{n=1}^{\infty} \begin{pmatrix} TPR(i+1)- \\ -TPR(i) \end{pmatrix} * \begin{pmatrix} FPR(i)+ \\ +FPR(i+1) \end{pmatrix} / 2, \qquad (7)$$

where $TPR(i)$ is the sensitivity (True Positive Rate) for the i-th threshold value; $FPR(i)$ – specificity (1 – False Positive Rate) for the i-th threshold value.

Owing to the calculation of the area under the curve, one can understand the measure of its "goodness", the further the curve is from the diagonal line, the better it is.

5. Results of determining the impact of data on work with video materials on the accuracy of models for predicting students' success

5. 1. Results of calculation of the prediction accuracy of models trained by the NB, LR, RF, and NN algorithms without data on students' work with video materials

The model's error matrix makes it possible to determine for how many students the prediction was made correctly. The general view of the obtained error matrices and models for the first data set is shown in Fig. 1.

Naive Bayes performs best in terms of specificity (highest True Negative is 73) but has relatively high false positive (84) and false negative (80) values, indicating potential problems with prediction accuracy. Logistic regression shows a better result in recognizing positive classes (True Positive – 301), but at the same time has a large number of false positives (102) results, which reduces its accuracy in predicting negative classes. Random Forest and Neural Networks showed the best sensitivity performance, with the lowest number of false negatives (24 and 15) and the highest True Positive (339 and 348), indicating their high ability to correctly identify positive cases.

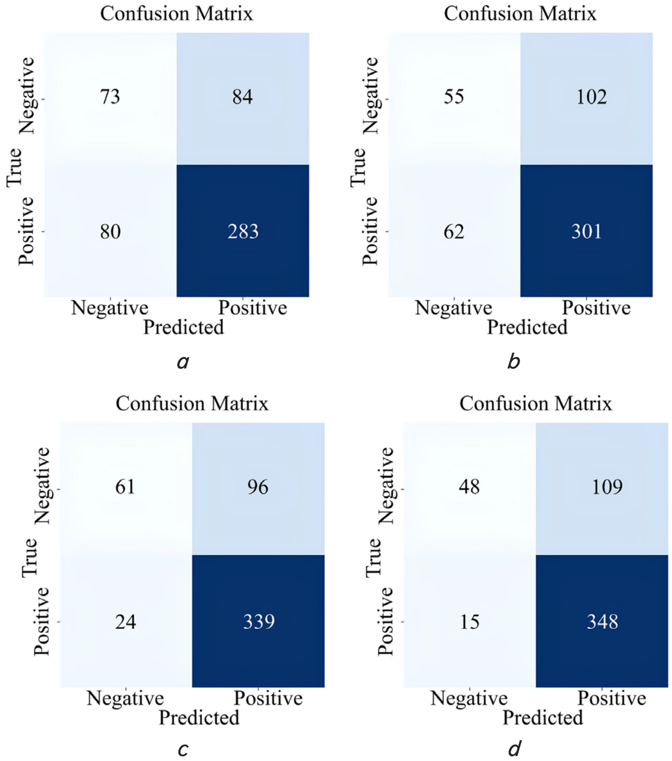

Fig. 1. Error matrices for constructed models with classifiers: *a* – naive Bayes; *b* – logistic regression; *c* – random forest; *d* – neural networks

Random forest and neural networks are the most effective in recognizing positive classes with the least number of errors on the current data. And naive Bayes and logistic regression have larger prediction errors.

Based on the obtained matrices, the values characterizing the overall accuracy of the classification and the effectiveness of the models were calculated, namely: overall accuracy, accuracy, balanced accuracy, sensitivity, specificity, *F*1*Score*, area under the curve (AUC), and ROC curve. The results of the calculations are given in Tables 3, 4.

Table 3

Calculations of values characterizing general accuracy and efficiency

Algorithm (classifier)	Overall accuracy	Accuracy	Sensitivity	Specificity
Naive Bayes	0.684	0.771	0.779	0.464
Logistic regression	0.684	0.746	0.829	0.350
Random forest	0.769	0.779	0.933	0.388
Neural networks	0.761	0.761	0.958	0.305

Table 4

Calculations of values characterizing general accuracy and efficiency

Algorithm (classifier)	Balanced accuracy	Area under the curve (AUC)	*F*1*Score*
Naive Bayes	0.622	0.607	0.775
Logistic regression	0.589	0.627	0.785
Random forest	0.661	0.738	0.849
Neural networks	0.632	0.681	0.848

Calculations showed that the random forest algorithm performs better in predicting success and has a higher accuracy of 76.9 %. The obtained accuracy value is only 0.8 % higher than that of the neural network algorithm. However, compared to the accuracy of naive Bayes and logistic regression, the increase in accuracy is already 8.5 %. The values of the True and False classes obtained during the calculation of Precision and *F*1*Score* are given in Table 5.

Among the considered classification algorithms, the best results are shown by the random forest, which demonstrated the highest overall accuracy (0.779) and balance between classes. This algorithm has high accuracy for the True class (0.78) and significantly better accuracy for the False class (0.72) compared to other models. The *F*1*Score* value for the True class (0.85) is also the highest, indicating a good ability to correctly recognize and not miss cases of this class. Although for the False class, this indicator is lower (0.50), which indicates some limitations in recognizing negative cases. Neural networks showed a high level of accuracy for both classes (True and False each 0.76), demonstrating balance and performance. However, the *F*1*Score* for the False class (0.44) indicates that the accuracy of these predictions is still problematic. Logistic regression and Naive Bayes had similar results with moderate overall accuracy (0.746 and 0.771, respectively) and poor ability to recognize the negative class (False), as evidenced by the low values of accuracy and *F*1*Score* for this class.

Based on the results, it can be concluded that the random forest is the most suitable for application under the conditions of this problem since it shows the best balance between the accuracy for both classes and the overall performance of the model. This algorithm provides the most stable predictions and minimizes the number of false positive cases, which makes it the optimal choice among the considered methods.

To evaluate the ability of the models to correctly classify, taking into account different values of the threshold value, ROC curves were constructed [23]. The ROC curve reflects the classifier's ability to correctly recognize positive classes and reject negative classes when the threshold value changes. It makes it possible to take into account the trade-off between the sensitivity and specificity of the classifier and make the examination of the results of the classification model more objective. The constructed plots of the ROC curve of the models for the first data set are shown in Fig. 2.

Table 5

True and False class values for Precision and *F*1*Score*

Algorithm (classifier)	Precision	True	False	*F*1*Score*	True	False
Naive Bayes	0.771	0.77	0.48	0.775	0.78	0.47
Logistic regression	0.746	0.75	0.47	0.785	0.79	0.40
Random forest	0.779	0.78	0.72	0.849	0.85	0.50
Neural networks	0.761	0.76	0.76	0.848	0.85	0.44

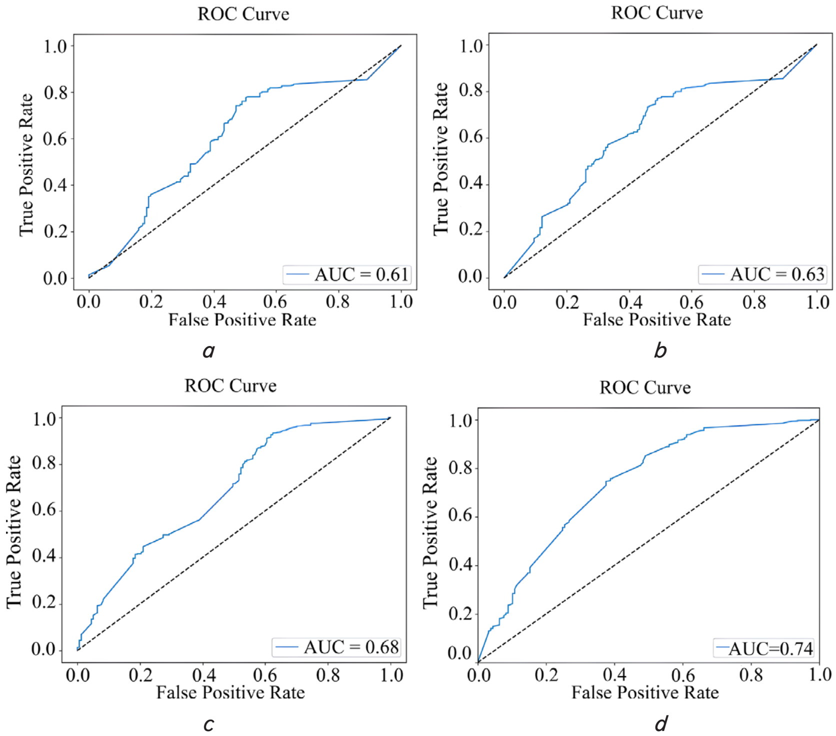

Fig. 2. Plots of ROC curves for models with algorithms: *a* – naive Bayes; *b* – logistic regression; *c* – neural networks; *d* – random forest

Constructed ROC-curves for the classification algorithms reflect the different level of ability of the models to distinguish between positive and negative cases. Random Forest has the highest area under the curve of 0.738, indicating its good overall classification performance. The curve for the random forest will be well above the diagonal line, showing high sensitivity (0.933) and precision (0.779), as well as a fairly high *F*1*Score* (0.849). This indicates a good ability of the model to detect positive cases, although the specificity (0.388) remains lower. Neural networks also show good results with an area under the curve of 0.681, indicating classification ability. The curve for neural networks will be located above the diagonal line, with high sensitivity (0.958) and precision (0.761), which provides a high value of *F*1*Score* (0.848). This shows high efficiency in detecting positive cases, but with moderate specificity (0.305). That can affect the accuracy of classification of negative cases. Logistic regression has an area under the curve (0.627), which is a moderate indicator of the effectiveness of the model. The logistic regression curve will lie slightly above the diagonal line, with high sensitivity (0.829), moderate specificity (0.350), and high *F*1*Score* (0.785). This indicates a good ability of the model to classify positive cases, but it has some difficulty in classifying negative cases accurately. Naive Bayes has the lowest area under the curve value of 0.607, indicating a moderate ability to distinguish classes. The curve for naive Bayes will be located closer to the diagonal line, showing good accuracy (0.771) and sensitivity (0.779), but with moderate specificity (0.464). The value of the *F*1*Score* for Naive Bayes (0.775) is the closest to the logistic regression results, indicating a relatively good combination of correct predictions and reduced errors.

Random forest and neural networks show the best results in distinguishing positive cases. Logistic regression also demonstrates good performance but with lower AUC and specificity values than random forest and neural networks. Naive Bayes yields the lowest results, indicating a relatively low classification ability of the model compared to other methods.

5. 2. Results of calculating the accuracy of models trained on data about students' work with video materials

The general view of the obtained error matrices and models for the second data set is shown in Fig. 3.

Based on our matrices, the values characterizing the overall classification accuracy were calculated, namely overall accuracy, accuracy, balanced accuracy, sensitivity, specificity, *F*1*Score*, area under the curve (AUC), and ROC curve. The results of the calculations are given in Tables 6, 7.

Calculations showed that the random forest algorithm performs better in predicting success on the input data and has a higher accuracy of 87.1 %.

However, compared to the accuracy of naive Bayes and logistic regression, the increase in accuracy is already 16.4 % and 10.6 %. The values of the True and False classes obtained during the calculation of Precision and *F*1*Score* are given in Table 8.

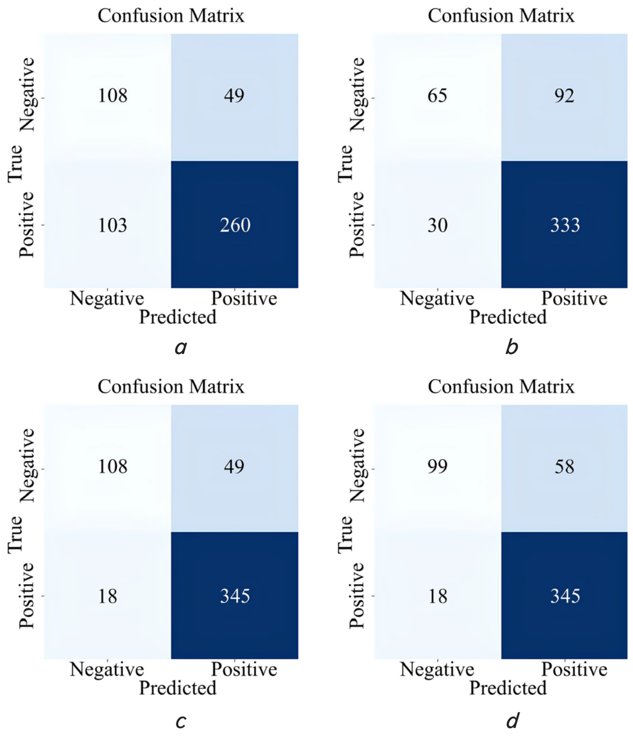

Fig. 3. Error matrices for constructed models with algorithms: *a* – naive Bayes; *b* – logistic regression; *c* – random forest; *d* – neural networks

Table 6

Calculations of values characterizing general accuracy and efficiency

Algorithm (classifier)	Overall accuracy	Accuracy	Sensitivity	Specificity
Naive Bayes	0.707	0.841	0.716	0.687
Logistic regression	0.765	0.783	0.917	0.414
Random forest	0.871	0.875	0.950	0.687
Neural networks	0.853	0.856	0.950	0.630

Table 7

Calculations of values characterizing general accuracy and efficiency

Algorithm (classifier)	Balanced accuracy	Area under the curve (AUC)	*F1Score*
Naive Bayes	0.702	0.776	0.773
Logistic regression	0.665	0.779	0.845
Random forest	0.819	0.875	0.911
Neural networks	0.790	0.859	0.900

Among the classification algorithms, the best results were demonstrated by the random forest, which has the best overall accuracy (0.875) and balance between classes. The random forest showed high accuracy for the True class (0.88) and fairly high accuracy for the False class (0.86). This provides a good balance between detecting positive and negative cases. The *F1Score* for the True class (0.911) and the False class (0.91) are the highest among all the considered algorithms. This shows the excellent ability of the model to correctly classify both positive and negative cases. Neural networks also performed well with accuracy for True class (0.86) and False class (0.85), but their *F1Score* for True (0.900) and False (0.90) class is slightly lower compared to Random Forest, although still high. From this we can conclude that neural networks do a good job of classifying both classes, but there is little difference in overall performance. Naive Bayes and logistic regression showed inferior results. Naive Bayes had an accuracy of 0.841, but a high False Positive Rate (0.51), indicating frequent errors in the classification of negative cases. Logistic regression also has moderate accuracy (0.783), but its results are less balanced. The low value of the *F1Score* for the False class (0.52) indicates difficulties in accurately classifying negative cases. Random Forest is the most efficient and balanced among the algorithms considered, providing the best accuracy and *F1Score* for both classes, making it the best choice for this task and the available data. To evaluate the ability of the models to correctly classify, considering different threshold values, ROC curves were constructed. The constructed plots of the ROC curves of the models for the second data set are shown in Fig. 4.

Constructed ROC curves for the considered classification algorithms demonstrate a high level of efficiency in distinguishing between positive and negative cases. The random forest achieves the highest area under the curve (0.875), indicating its excellent ability to classify both classes. The curve for the random forest is located well above the diagonal line, indicating high accuracy (0.875) and sensitivity (0.950). A high *F1Score* (0.911) indicates strong performance in recognizing both positive and negative cases. Neural networks also demonstrate a high level of performance with an area under the curve (0.859) showing good classification ability. The curve for neural networks is located above the diagonal line, with high sensitivity (0.950) and precision (0.856), which provides a high *F1Score* (0.900). This indicates that the model is effective in recognizing positive cases and relatively good in distinguishing negative ones. Logistic regression has an area under the curve (0.779), a value smaller than that of random forest and neural networks, but still shows good classification ability. The curve for the logistic regression will lie above the diagonal line, with high sensitivity (0.917) and moderate specificity (0.414), and an *F1Score* (0.845), indicating good balance, especially for positive cases. Naive Bayes has the lowest area under the curve (0.776), indicating moderate classification ability. The curve for naive Bayes will lie closer to the diagonal line, showing good accuracy (0.841) and sensitivity (0.716), but with less high specificity (0.687). The *F1Score* for Naive Bayes (0.773) is the lowest among the considered algorithms, indicating a lower efficiency in combining correct predictions and error minimization.

Table 8

True and False class values for Precision and *F1Score*

Algorithm (classifier)	Precision	True	False	*F1Score*	True	False
Naive Bayes	0.841	0.84	0.51	0.773	0.77	0.59
Logistic regression	0.783	0.78	0.68	0.845	0.85	0.52
Random forest	0.875	0.88	0.86	0.911	0.91	0.76
Neural networks	0.856	0.86	0.85	0.900	0.90	0.72

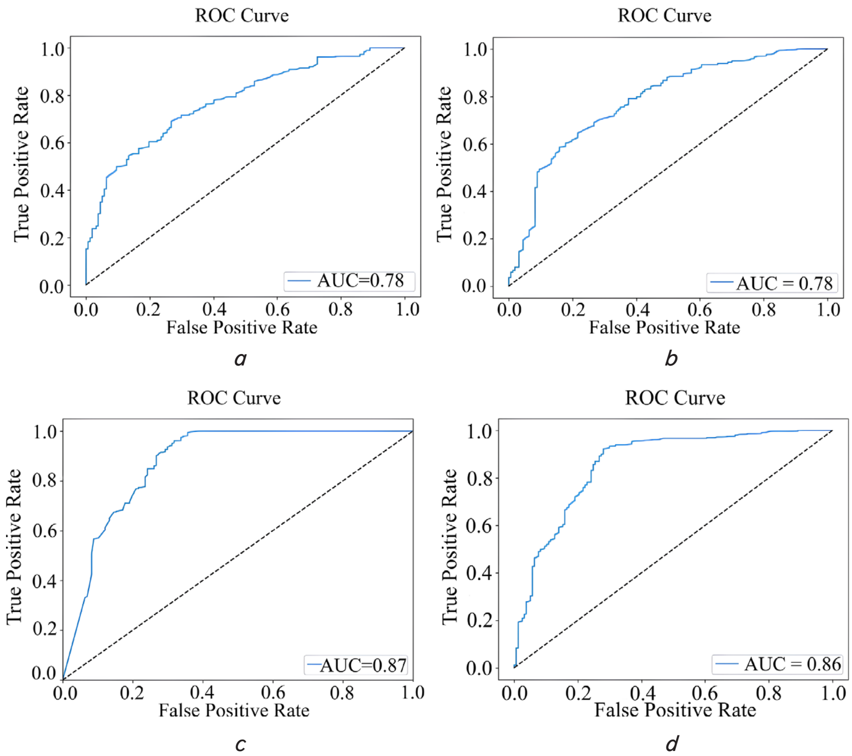

Fig. 4. Plots of ROC curves for models with algorithms: *a* — naive Bayes; *b* — logistic regression; *c* — random forest; *d* — neural networks

Random Forest and Neural Networks show the best overall performance, with high AUC and *F*1*Score* values, while Logistic Regression also performs well, but with less performance compared to Random Forest. Naive Bayes has the lowest results among all models, especially in the combination of accuracy and sensitivity. This result is a fairly good starting point, but in the process of further research, additional refinement may be needed to improve these indicators. Random forest is an ensemble method that usually works better in cases where the relationships between features and the original classes are more complex, non-linear, or when there are many features. It can automatically consider feature importance and make better predictions than linear models on complex data.

5. 3. Calculation of the influence of information about students' work with video materials on the accuracy of models for predicting students' success

The results of the obtained increase in values characterizing the accuracy of forecasting models between the first and second data sets are given in Tables 9, 10.

Table 9

Result of the obtained increase in forecasting accuracy

Algorithm (classifier)	Overall accuracy	Accuracy	Sensitivity	Specificity
Naive Bayes	+2.3 %	+7 %	–6.3 %	+22.3 %
Logistic regression	+8.1 %	+3.7 %	+8.8 %	+6.4 %
Random forest	+10.2 %	+9.6 %	+1.7 %	+29.9 %
Neural networks	+9.2 %	+9.5 %	–0.8 %	+32.5 %

Table 10

Result of the obtained increase in forecasting accuracy

Algorithm (classifier)	Balanced accuracy	Area under the curve (AUC)	*F*1*Score*
Naive Bayes	+8 %	+16.9 %	–0.002
Logistic regression	+7.6 %	+15.2 %	+0.006
Random forest	+15.8 %	+13.7 %	+0.062
Neural networks	+15.8 %	+17.8 %	+0.052

The highest increase in accuracy with a difference of 1.8 % was shown by models with algorithms: random forest – 87.1 % and neural networks – 85.3 %. The accuracy gain was almost 10 %, the balanced accuracy increased by 15 %, and the overall efficiency expressed by the area under the curve (AUC) increased by 14 %. The forecasting accuracy of models with naive Bayes and logistic regression algorithms was 70.7 % and 76.5 %, and the increase was 2.3 % and 8.1 %, respectively. The result shows that the addition of data on interaction with video materials has a good effect on increasing the accuracy of prediction.

6. Discussion of results based on determining the impact of data on work with video materials on the accuracy of models for predicting students' success

With the help of state-of-the-art technologies and machine learning, problems that have to be faced every day are solved. One of them is predicting the success of students studying at universities. In this work, it is recommended to predict success using the following features: attendance, evaluations, and interaction with educational video materials. Data on interaction with video materials were obtained from the UVPlayer plugin integrated into the Moodle system. Two sets of data were used to compare the increase in accuracy of forecasting models by adding data on interaction with video materials (Tables 1, 2). The first set contained only data on attendance and grades. In the second set, data on interaction with educational video materials was added. For each of the sets, models were built to predict success with algorithms: logistic regression, naive Bayes, random forest, and neural networks. And the calculated characteristics responsible for the accuracy of forecasting. For the first set of data, the calculations are given in Tables 3, 4; for the second set of data, the calculations are given in Tables 6, 7. The values of the classes obtained during the calculation of accuracy and *F1Score* are given in Table 5 for the first set of data, and in Table 8 for the second data set. Error matrices for the created prediction models based on the first data set are shown in Fig. 1, and the constructed plots of ROC curves are shown in Fig. 2. Error matrices for the created forecasting models based on the second data set are shown in Fig. 3, and the constructed plots of ROC curves are shown in Fig. 4. Our results confirmed the hypothesis that additional data increases the accuracy of forecasting. The determination of the obtained increase in the accuracy of predicting success among models with algorithms of logistic regression, naive Bayes, random forest, and neural networks between two data sets is given in Tables 9, 10. The increase in accuracy was more than 10 %, the balanced accuracy increased by 15 %, and the overall efficiency expressed by the area under the curve (AUC) increased by 14 %. The study showed that models with random forest and neural network algorithms give better prediction accuracy of 87.1 % and 85.3 % on the available data compared to naive Bayes and logistic regression algorithms whose accuracy was 70.7 % and 76.5 %, respectively. The smallest increase in prediction accuracy is 2.3 % in the model with the naive Bayes algorithm. The choice between random forest and neural networks often depends on the specific requirements of the problem. If simplicity and speed of implementation are important, and the simulation is not too complex, a random forest may be the best choice. If you need to deal with large amounts of data and complex dependences, neural networks can be more efficient, although they will require more effort and resources.

In contrast to [10], in which success prediction was performed on the basis of grades received by students and [11], in which success prediction was performed on the basis of student visits, the obtained accuracy of the current model with video interaction data is 10 % higher. The disadvantage of [10] is that only grades are taken as a feature for classification, so it may be difficult to get high accuracy when predicting success at an early stage. The disadvantage of [11] is that although attendance is an indicator that directly affects success, it cannot be the only sufficient factor. Even with 100 % attendance, a student may have low scores and fail to pass the session test. The obtained prediction accuracy, as in [11], is less than 75 %. Unlike previous works [12], the number of viewed videos was also used to predict success, while the accuracy was higher and amounted to 96.3 %. However, it is not clear how exactly the number of viewed videos is calculated. Is there a check that the video is actually watched to the end and not just opened by the student? The model shows high accuracy, but it is not clear what the prediction result would be at an early stage. Unlike previous work [13], success prediction was performed on the activity log database, neural networks achieved the highest classification accuracy of 90 %. The disadvantage is that most of the presented data from the activity log about users does not convey information about the actual interaction of the user with them. And only the very fact of switching to an activity or opening it, for example, viewing a task. But a review is not proof of an attempt to solve a specific task. In the current work, the data on interaction with video materials, on the contrary, clearly show the aspects of the user's interaction with the video during the viewing process. In contrast to previous works, in [14] the following features were used to predict success: email communication, joint content creation using a wiki, content interaction measured by file views, and self-assessment using online tests. As a result, the model predicted success with a high accuracy of 98.3 %. The obtained accuracy is also confirmed by the built ROC curve and the value of the calculated area under the curve (AUC) is 98.9 %. This highlights the fact that data relevant to the actual interaction can significantly improve the resulting model prediction accuracy. In work [15], as well as in [10, 11], student evaluation data were used to predict student success. But unlike previous studies, this one took into account data from 3 courses, which contributes to more accurate prediction of the model. This suggests that the more data for a longer period of time, the more accurate the model can be. Work [16] gives the results of research on the development of accurate prediction of the results of the course completion by students. Unlike previous studies, this study considered demographic, assessment, and student interaction data to provide comprehensive predictions. The obtained accuracy of classification of models for 2 classes was more than 90 % and more than 70 % for four classes. This predictive approach provides insight into course student outcomes, offering valuable information to improve student engagement in online learning environments.

A distinctive feature of this work compared to [10–16] is the use of additional data on actual interaction with educational video materials. The very fact of opening an educational resource, for example viewing a lecture or an exercise, is not proof of an attempt to solve a specific task.

The limitation of our research is that students' work with educational materials should be carried out mainly within the framework of LMS and recorded in the database. That is, there should be a direct relationship between the intensity of students' work and the number of their actions in the LMS.

The disadvantages of the study are:

– a lack of the possibility to take into account specific features of disciplines;

– a lack of the possibility to take into account the basic level of students' knowledge.

In both cases, approximately the same performance indicators of students in LMS can correspond to significantly different results of knowledge assessment, which will reduce the accuracy of the models.

Further development of the research involves reducing the impact of shortcomings and limitations, namely:

– expansion of data sets for model training through the use of information on intermediate results of knowledge assessment;

– introduction of parameters characterizing the features of the disciplines into the data for model training;

– checking the accuracy of predicting success using other types of models.

It is also planned to improve the developed UVPlayer plugin for the Moodle system, in particular, to save more detailed information about students' work with video materials.

7. Conclusions

1. Our calculations showed that the random forest algorithm performs better in predicting success and has a higher accuracy of 76.9 %. However, compared to the accuracy of naive Bayes and logistic regression, the increase in accuracy is already 8.5 %.

2. The calculations demonstrated that the random forest algorithm performs better in predicting success on the input data and has a higher accuracy of 87.1 %. However, compared to the accuracy of naive Bayes and logistic regression, the increase in accuracy is already 16.4 % and 10.6 %.

3. The highest increase in accuracy with a difference of 1.8 % was shown by models with algorithms: random forest – 87.1 % and neural networks – 85.3 %. Accuracy increased by 10 %, balanced accuracy increased by 15 %, and overall efficacy expressed as area under the curve (AUC) increased by 14 %. Whereas the forecasting accuracy of models with naive Bayes and logistic regression algorithms was 70.7 % and 76.5 %, and the increase was 2.3 % and 8.1 %, respectively. Our result shows that the addition of data on interaction with video materials has a good effect on increasing the accuracy of prediction.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

The data will be provided upon reasonable request.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

References

1. Liu, M., Yu, D. (2022). Towards intelligent E-learning systems. Education and Information Technologies, 28 (7), 7845–7876. https://doi.org/10.1007/s10639-022-11479-6
2. Soloshych, I., Grynova, M., Kononets, N., Shvedchykova, I., Bunetska, I. (2021). Competence and Resource-Oriented Approaches to the Development of Digital Educational Resources. 2021 IEEE International Conference on Modern Electrical and Energy Systems (MEES), 2, 1–5. https://doi.org/10.1109/mees52427.2021.9598603
3. Panasiuk, O., Akimova, L., Kuznietsova, O., Panasiuk, I. (2021). Virtual Laboratories for Engineering Education. 2021 11th International Conference on Advanced Computer Information Technologies (ACIT), 1, 637–641. https://doi.org/10.1109/acit52158.2021.9548567
4. Bradley, V. M. (2020). Learning Management System (LMS) Use with Online Instruction. International Journal of Technology in Education, 4 (1), 68. https://doi.org/10.46328/ijte.36
5. Gamage, S. H. P. W., Ayres, J. R., Behrend, M. B. (2022). A systematic review on trends in using Moodle for teaching and learning. International Journal of STEM Education, 9 (1). https://doi.org/10.1186/s40594-021-00323-x
6. Bognár, L., Fauszt, T. (2022). Factors and conditions that affect the goodness of machine learning models for predicting the success of learning. Computers and Education: Artificial Intelligence, 3, 100100. https://doi.org/10.1016/j.caeai.2022.100100
7. Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., Durán-Domínguez, A. (2020). Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. Applied Sciences, 10 (3), 1042. https://doi.org/10.3390/app10031042
8. Sáiz-Manzanares, M. C., Marticorena-Sánchez, R., García-Osorio, C. I. (2020). Monitoring Students at the University: Design and Application of a Moodle Plugin. Applied Sciences, 10 (10), 3469. https://doi.org/10.3390/app10103469
9. Gaftandzhieva, S., Talukder, A., Gohain, N., Hussain, S., Theodorou, P., Salal, Y. K., Doneva, R. (2022). Exploring Online Activities to Predict the Final Grade of Student. Mathematics, 10 (20), 3758. https://doi.org/10.3390/math10203758
10. Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9 (1). https://doi.org/10.1186/s40561-022-00192-z
11. Pylypenko, V., Statsenko, V. (2024). Assessment of the efficiency of the success prediction model using machine learning methods. Herald of Khmelnytskyi National University. Technical Sciences, 1 (3 (335)), 349–356. https://doi.org/10.31891/2307-5732-2024-335-3-47
12. Ljubobratović, D., Matetić, M. (2020). Using LMS activity logs to predict student failure with random forest algorithm. INFuture2019: Knowledge in the Digital Age. https://doi.org/10.17234/infuture.2019.14
13. Aleksandrova, Y. (2019). Predicting students performance in moodle platforms using machine learning algorithms. Conferences of the department Informatics, 1, 177–187. Available at: https://informatics.ue-varna.bg/conference19/Conf.proceedings_Informatics-50.years%20177-187.pdf
14. Zacharis, N. (2016). Predicting Student Academic Performance in Blended Learning Using Artificial Neural Networks. International Journal of Artificial Intelligence & Applications, 7 (5), 17–29. https://doi.org/10.5121/ijaia.2016.7502

15. Tamada, M. M., Giusti, R., Netto, J. F. de M. (2022). Predicting Students at Risk of Dropout in Technical Course Using LMS Logs. Electronics, 11 (3), 468. https://doi.org/10.3390/electronics11030468
16. Althibyani, H. A. (2024). Predicting student success in MOOCs: a comprehensive analysis using machine learning models. PeerJ Computer Science, 10, e2221. https://doi.org/10.7717/peerj-cs.2221
17. Jabbar, H. K., Khan, R. Z. (2014). Methods to Avoid Over-Fitting and Under-Fitting in Supervised Machine Learning (Comparative Study). Computer Science, Communication and Instrumentation Devices, 163–172. https://doi.org/10.3850/978-981-09-5247-1_017
18. Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.
19. Sial, A. H., Rashdi, S. Y. S., Khan, A. H. (2021). Comparative Analysis of Data Visualization Libraries Matplotlib and Seaborn in Python. International Journal of Advanced Trends in Computer Science and Engineering, 10 (1), 277–281. https://doi.org/10.30534/ijatcse/2021/391012021
20. Sapre, A., Vartak, S. (2020). Scientific Computing and Data Analysis using NumPy and Pandas. International Research Journal of Engineering and Technology, 7, 1334–1346.
21. Krstinić, D., Braović, M., Šerić, L., Božić-Štulić, D. (2020). Multi-label Classifier Performance Evaluation with Confusion Matrix. Computer Science & Information Technology. https://doi.org/10.5121/csit.2020.100801
22. Lavazza, L., Morasca, S., Rotoloni, G. (2023). On the Reliability of the Area Under the ROC Curve in Empirical Software Engineering. Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering. https://doi.org/10.1145/3593434.3593456
23. Bowers, A. J., Zhou, X. (2019). Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes. Journal of Education for Students Placed at Risk (JESPAR), 24 (1), 20–46. https://doi.org/10.1080/10824669.2018.1523734