

Article

# Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques

Janka Kabathova  and Martin Drlik \* 

Department of Informatics, Faculty of Natural Sciences, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 01 Nitra, Slovakia; janka.kabathova@ukf.sk

\* Correspondence: mdrlik@ukf.sk

**Featured Application:** The found model with the best values of the performance metrics, found as the result of comparing several machine learning classifiers, can identify students at risk even though the set of educational data obtained during the course run is scarce. As a result, a suitable form of intervention at the individual e-learning course level can be applied in time.

**Abstract:** Early and precisely predicting the students' dropout based on available educational data belongs to the widespread research topic of the learning analytics research field. Despite the amount of already realized research, the progress is not significant and persists on all educational data levels. Even though various features have already been researched, there is still an open question, which features can be considered appropriate for different machine learning classifiers applied to the typical scarce set of educational data at the e-learning course level. Therefore, the main goal of the research is to emphasize the importance of the data understanding, data gathering phase, stress the limitations of the available datasets of educational data, compare the performance of several machine learning classifiers, and show that also a limited set of features, which are available for teachers in the e-learning course, can predict student's dropout with sufficient accuracy if the performance metrics are thoroughly considered. The data collected from four academic years were analyzed. The features selected in this study proved to be applicable in predicting course completers and non-completers. The prediction accuracy varied between 77 and 93% on unseen data from the next academic year. In addition to the frequently used performance metrics, the comparison of machine learning classifiers homogeneity was analyzed to overcome the impact of the limited size of the dataset on obtained high values of performance metrics. The results showed that several machine learning algorithms could be successfully applied to a scarce dataset of educational data. Simultaneously, classification performance metrics should be thoroughly considered before deciding to deploy the best performance classification model to predict potential dropout cases and design beneficial intervention mechanisms.

**Keywords:** learning analytics; educational data mining; machine learning; dropout prediction



**Citation:** Kabathova, J.; Drlik, M. Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques. *Appl. Sci.* **2021**, *11*, 3130. <https://doi.org/10.3390/app11073130>

Academic Editor: Juan A. Gómez-Pulido

Received: 10 February 2021

Accepted: 30 March 2021

Published: 1 April 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Nowadays, Virtual Learning Environments (VLEs) represent mature online education platforms allowing teachers to provide courses, which contain thoroughly managed resources and myriads of different engaging activities. VLEs allow creating various educational games, quizzes, e-learning tests, or prepare multimedia materials to attract students.

Considering technological progress and increasing demand for online education, VLEs have perceived fast development over the past few years [1]. VLE developers and providers have fully taken advantage of the internet and current web technologies and operate modern education systems to improve students' knowledge and skills. Educational approaches and methodologies tailored to the functionalities of the virtual platforms get rid of the limitations of traditional courses taken in classrooms and come up with higher flexibility in terms of where and when to take the courses online.

Modern computational tools for data analysis, standardization of formats for storing data, and increasing computing power allow research questions to be explored widely using diverse data-driven approaches. Since virtual learning environments are designed to capture and store as much data on user interaction with the system as possible, it is not surprising that this data is analyzed using different knowledge discovery techniques to identify the learning-related problems early and suggest the most suitable form of intervention. As a result, different educational data mining and learning analytics techniques have already helped analyze and partially understand students' behavior, detect frequent learning mistakes, recommend learning paths, improve tests, and measure students' performance.

Despite the rapid development and successes of data-driven research, educational institutions still face many principal difficulties such as effective VLE stakeholders' interaction and immediate feedback, which are not appropriate, insufficient, or delayed. This fact subsequently influences student's educational performance and their final outcome [2].

Student's dropout poses one such significant problem within contemporary e-learning approaches [3]. The dropout has a negative impact on educational organizations as well as stakeholders. Even though it has been the focus of many research, e-learning dropout rates generally tend to be higher than face-to-face education [4].

Therefore, the researchers thoroughly apply different knowledge discovery methods and examine the student's behavior, their characteristic features, as well as the impact of the external environment on the learning process and outcomes to minimize the dropout, create effective intervention strategy and provide immediate feedback [5].

Educational data used in dropout prediction models comes from different resources, has different size and quality. Educational data is often stored in database systems. It can consist of information about students' activities and provide end-users with course-wide statistics, giving an aggregated overview of all stakeholders' activity. Moreover, summary information about partial and final grades are also arranged about each student's performance. Finally, event logs are often used, which indirectly track every stakeholder's interaction with the platform [6].

The availability of the educational data resources for data-driven research differs depending on the researcher's privileges and the scope of the research problem. Whereas the researcher can have a considerable amount of data available in the form of logs, the amount of the data describing the students' achievements at the course level is limited. Combining these resources does not automatically correlate with the higher accuracy of the final predictions of student's dropout because the data must be often aggregated.

Therefore, the main goal of the article is to emphasize the importance of the data understanding, data gathering phase, stress the limitations of the available datasets of educational data, compare the performance of several machine learning classifiers, and show that also a limited set of features, which are available for teachers in the e-learning course, can predict student's dropout with sufficient accuracy if the performance metrics are thoroughly considered.

Even though the main topic of the paper is quite frequent, the paper describes a case study, which emphasizes the necessity to follow well-known knowledge discovery methodology even in the case of commonly used classifiers and consider not only a limited set of performance metrics. The applied approach has the potential to predict the student's dropout at the course level if the course structure has not significantly changed and the course has been re-opened for several years.

The research paper is organized as follows. At first, the summary of related works focused on effective predictions of students' success using machine learning techniques in the learning analytics domain is presented. The individual phases of the CRISP-DM methodology, which manages the knowledge discovery process, are gradually introduced, starting Section 3. This section provides information about the main characteristics of the researched dataset. Subsequently, data understanding and pre-processing phases are described in detail. Section 4 deals with the models of prediction of the student's dropout, in which different machine learning classifiers had been applied. Section 5

concerns the obtained results and evaluation of the models based on different performance measures. Moreover, the performance of the machine learning classifiers is compared using McNemar's test. Finally, Section 6 summarizes the research results, discusses the research limitation, and proposes its possible future direction.

## 2. Related Works

Various definitions of students' dropout were established in previous studies. The most frequent definition focuses on whether students will still be active until the last week or whether the current week is the final week in which students are active. Early identification of students at risk of dropout plays an important role in reducing the issue and allowing required conditions to be targeted.

Timing considerations are, therefore, significant for the dropout problem. As a result, some analyses have shown that 75% of dropouts occur in the first weeks [7]. The prediction of dropout is often considered a time series prediction problem or a sequence labelling problem [8]. These can correspond with the final status of students [9]. On the other hand, the time dimension can be included in prediction dropout indirectly using the input features available in a particular time window, which allow selecting a suitable form of an intervention [10].

Prediction of the student's learning success or failure belongs to one of the most frequently researched topics of educational data mining (EDM) and learning analytics (LA) disciplines [3]. EDM deals with the analysis of study-related data to understand student behavior. These techniques are usually applied to provide more effective learning environments through revealing useful information for modifying the structure of courses or support the prediction of student performance and behavior [11]. On the other hand, learning analytics deals with the measurement, collection, analysis, and reporting of student data and their backgrounds in order to understand and improve learning and the environments in which it takes place [12]. The core of current prediction approaches is usually based on EDM and LA methods. One of the hottest research topics in learning analytics has been predicting the probability of students completing or failing a course, especially from the very early weeks [13], as well as EDM [14]. Once a reliable performance prediction is available, it can be used in many contexts to identify weak students and provide feedback to students or predict students' failure [15].

A successful prediction depends on the use of reliable, accurate, and significant data [9]. The size of the dataset is a typical pitfall of many research studies published in educational technology, learning analytics and educational data mining research [16]. It is challenging to reach the volume of educational data, which would fully satisfy the machine learning requirements besides using logs [8]. On the other hand, the high granularity and technical nature of the VLE logs often miss the necessary semantics, which can cause problems, how the findings should be interpreted from the educational point of view [17].

Hämäläinen and Vinni indicated in their review of classifiers for EDM that the datasets often used in educational data mining are relatively small. They usually consist of 200 records on average on the course level [18]. Many analyzed experiments confirmed this statement. A similar data volume was used to evaluate the machine learning techniques on the individual VLE course level [18].

As Ifenthaler and Yau [16] stated, more educational data has not always led to better educational decisions. The same authors concluded that the reliability and validity of data and its accurate and bias-free analysis are critical for generating useful summative, real-time, or formative, and predictive or prescriptive insights for learning and teaching.

Lang et al. [19] stated that classification and regression methods, neural network, Bayesian networks, support vector machines, logistic regression, and linear regression could solve the student performance prediction problem. These models are often called black-box models that are difficult to understand and interpret. All of them also rely heavily on feature extraction. Feature extraction refers to the compilation of a subset of unique predictive features for the predictive problem in modelling, also known as attribute

selection. The process helps to identify relevant attributes in the dataset that contribute to the accuracy of the prediction model, for instance, by the last activity on the student's corresponding course to predict a drop out [4]. Strategies effective for one kind of datasets may not be sufficient for another one. In this case, it is often necessary to manually develop new strategies for extracting features [6].

On the other hand, decision trees, rule-based classifier belong to the white-box models that are more understandable and easily interpretable because they expose the reasoning process underlying the predictions. Another possibility is to use clustering algorithms or association rule mining. Correlation analysis of grades achieved from courses and attributes defining credits obtained by students and their average grades can also be valuable [19].

Ifenthaler et al. [16] realized a review of contemporary LA research direction. They distinguish different sets of predictors of students' dropout. The first set of predictors is based on online behavior data, mainly log files and trace data. In this context, one of the difficult challenges is to determine how to engage students to participate in online learning [20]. The second set of predictors is based on students' background information, such as demographics, socioeconomic status, prior academic experience, and performance [7]. Many studies usually also examine student-related records, e.g., the age, gender, status, because they are readily available in the information systems of each university. It is proven that attributes like students' average grade or average grade obtained in a course positively affect prediction [21].

Analysis of students' logs together with some other predictors, such as the number of discussion posts, number of forum views, number of quiz views, number of module views, number of active days, and social network activities, were aimed to predict the dropout through the implementation of a decision tree [22].

Similarly to the research presented in this paper, Kennedy et al. [23] considered the popular types of variables, such as task submissions, assignment switches (i.e., the amount of times a learner goes from one assignment to another), active days, and total points. They concluded that these features could predict the students' success. However, they did not discuss the role of the smaller dataset and performance metrics in detail.

Skalka et al. noted that other special activities, like programming assignments, can also be considered to predict the final score in the course effectively [12]. The assignments have also been analyzed in [24,25]. Both studies confirmed that assignments could be successfully used for prediction students' learning outcomes.

Considering the role of forum posts as predictors is also a significant subject of research. They can be an appropriate source of predictive elements (e.g., the number of posts, number of comments, number of votes up, and so on) that are not normally available in face-to-face courses [1].

Herodotou et al. [26] demonstrated that the courses in which teachers used predictive methods delivered at least 15% higher output than the classes without their use [1]. Nevertheless, it is challenging to reach this state because the fewer data is available, the less accurate prediction can be obtained.

Moreover, researchers are trying to find extra features that can contribute to a deeper understanding of students. Overall, all these approaches produced acceptable results, allowing the prediction of students' performance to achieve overall accuracy varying from 65% to 90% [4]. It is a considerable challenge to optimize the results obtained by predictions because different algorithms typically present a large variance in performance rates depending on the combination of several features (e.g., balance among classes, amount of data, input variables, and others) [1].

### 3. Data Preparation

The purpose of the knowledge discovery process is to extract and transform the information into an understandable structure for further use. Data mining is a part of this process, which involves integrating techniques such as statistics, machine learning,

pattern recognition, data visualization, or data analysis [27]. The process is often iterative and focused on discovering valuable relationships and information by analyzing many variables.

The knowledge discovery process has been continually improved and transformed into several practical variations. One of the most widely used is the Cross-Industry Standard Process for Data Mining, known as CRISP-DM. As the creation of predictive models in this paper is a typical task of a knowledge discovery problem, it was also driven by the CRISP-DM methodology [17].

### 3.1. Business Understanding

This initial phase of CRISP-DM focuses on understanding the problem and requirements from a broader perspective. It defines the goals of analysis and converts them into a data mining problem definition and plan of achieving the goals.

The data used in this research comes from the e-learning course of LMS Moodle, which has been taught for four academic years. The course contained many resources and interactive activities, which engage students to be more active and receive knowledge and skills, mainly based on learning.

The natural question is how the points obtained from these activities and the students' behavior during the term can contribute to the effective prediction of students' success or failure.

There are many more or less data-driven approaches, which can provide partial answers. The more conventional pedagogical research assumes the implementation of precisely defined phases, including pre-test, post-test, and hypotheses evaluation to identify suitable predictors. The size of the dataset is considered restrictive in this case but compensated by the appropriately chosen statistical tests. Limited options for the automation of the individual steps of this process hinder its wider use by LMS stakeholders.

Therefore, contemporary educational research puts hope to the data-driven approaches, which promise partial automation of the predictive process, potentially more straightforward interpretation, as well as an effective selection of suitable kind of intervention for a wider group of stakeholders.

Even though the LMS Moodle has already implemented its core own learning analytics model based on the students' behavior in different course activities, it does not consider students' partial outcomes, and the stakeholders do not often use it.

Therefore, different machine learning algorithms were applied in this case study to obtain accurate prediction as there is still no specific way to present the best solution for predicting students' performance in online education or identifying students who are most likely at risk of failing the course.

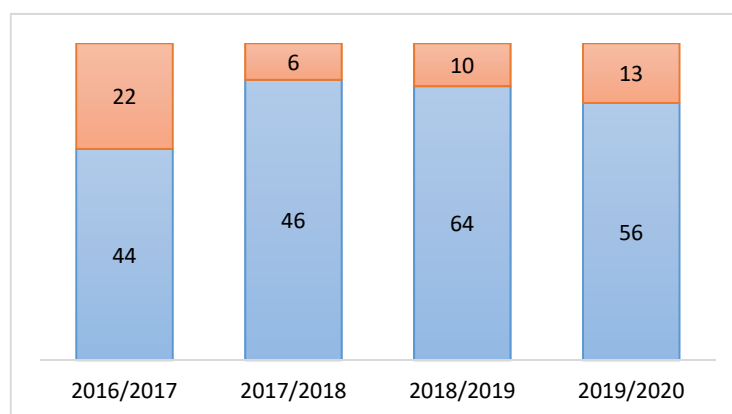
### 3.2. Data Understanding

The second stage involves acquiring the information or access to the data required to solve the problem. This initial collection includes procedures for data retrieval, description of data, and verification of the quality and accuracy of the data [27].

Educational data extracted from VLEs databases can be used for finding key attributes or dependencies among courses. Each VLE offers data suitable for knowledge discovery tasks. The knowledge obtained from logs and study results can enhance student's profiles [28]. The more is known about students' performance during the study course, the better and more useful prediction model can be created.

The step of data understanding is essential to discover useful data patterns successfully [25]. The educational data used in this paper comes from 261 unique students enrolled in the introductory course of Database Systems at the Constantine the Philosopher University in Nitra between the years 2016 and 2020. Data has been anonymized during the extraction process.

Subsequently, the proportion of dropout students was determined in each academic year. As a result, 21.92% of students in total failed to pass the course, as shown in Figure 1.



**Figure 1.** Distribution of students' dropout over several academic years (orange—number of failed students, blue—number of students who successfully finished the course).

The reasons for students' dropout were different. The insufficient number of obtained points, delayed or non-submitted assignment, as well as low activity in the course represent the most frequent reasons. However, a particular dropout reason has not been analyzed and has not played any role in the ongoing research.

One of the already mentioned LA and EDM open questions is how to effectively predict student performance based on their activity and partial outcomes during the study. Student performance was examined according to available attributes in the course. The aim was to identify the extent to which the received final grades corresponding to the unsuccessful completion of the course could be explained or predicted by given attributes from the course.

The course was divided into 13 sections according to the university standards. Each section contained several educational resources, as well as more than three assignments in the form of different LMS Moodle activities and quizzes. The assignments were used predominantly in the first two thirds of the term, followed by the midterm test and project assignment.

### 3.3. Data Cleaning and Feature Selection

Data preparation involves cleaning of data, data joining, attributes transforming, and feature selection or reduction. Various methods can be applied to analyze the data based on the data understanding in order to uncover hidden patterns [27]. The output from this phase is a construction of the final dataset deemed to be imported from the initial raw data into the modelling tool.

The educational data is often either sparse, noisy, inconsistent, or include several attributes that are irrelevant. Therefore, this step is usually the one where the aim is to remove all unwanted data characteristics. Noisy data can be smoothed by binning, regression, or clustering. There are also several techniques on how to overcome the problem of missing values. The incomplete records can be removed or imputed with a global constant or the most probable value [9]. However, the semantics of the missing values should be taken into account in all cases individually.

As was already mentioned, the educational data used in this paper represents 261 unique students enrolled in the e-learning course used in blended study form over four academic years. The results of interactive learning activities were exported using a conventional tool of LMS Moodle. Therefore, it was not required to identify individual students. The personal data have been removed.

Simultaneously, the detailed records about their activity in the form of application logs were also exported. However, this data had to be cleaned in contrast to the grades received by the students. The reason is that these records contained information about other stakeholders like teachers and administrators, as well as information about activities,

which did not relate directly to the learning process. After cleaning, the records were aggregated for each student and joined with his partial outcomes.

The number of records represents a persistent issue in the educational technology, learning analytics and educational data mining research field [18]. This problem can be overcome if the research methods use only the logs, representing the digital traces of the VLE stakeholders' behavior. On the other hand, if the input dataset should also contain the students' achievements or grades, the number of unique records is radically reduced, and the records should be aggregated [29]. The reason is that data about the students' partial achievements should be often aggregated to minimize the problem with data incompleteness, presence of missing values or voluntary character of the assignments, which the student could submit. The missing values cannot be easily removed because of their semantics. It may not have been known if the students submitted the assignment and received zero points, did not participate in the activity, or selected other voluntary assignment. Simultaneously, this record cannot be removed from the already limited size of the dataset.

Considering research reviews in the learning analytics and educational data mining domain, the dataset used in this case study can be considered sufficient [18]. Simultaneously, it is necessary to take into account the reduction of results interpretation and more limited options to generalize the obtained results when machine learning techniques are applied for such scarce datasets [30].

Precise and robust prediction models depend on thoroughly selected features extracted from the studied dataset [8]. Therefore, the feature extraction process is critical as it directly impacts the performance of any prediction model. According to previous works, different studies often use different features [7]. However, they all generally refer to the students' interactions with the course content and the traces of their activities on the VLE platform. They represent the natural behavioral features that the LMS Moodle platform keeps track of [31,32].

As was mentioned earlier, the course used as the primary source of data contained 13 sections with more than 30 interactive activities, which required continual students' activity. These activities could be divided into assignments, tests, project, and exam. The partial results obtained from these activities and the number of accesses to the course could be considered an ideal input set of features for machine learning algorithms at first glance.

Therefore, they were intended to process using standard feature selection or dimensionality reduction technique. Consequently, different levels of features granularity were considered, from the individual activity levels, over different kinds of their categorization or aggregation with the aim to apply some feature selection technique.

Although the importance of this technique increases with the number of the features in the input dataset, the restricted number of features and records influenced the final decision not to use the automated feature selection technique and replace it with the aggregation based on the category of activity. In other words, even though the partial points obtained from the individual activities could be considered in the input dataset as individual features, they were replaced by the total sum for the category of activity in the observed period of the term. As a result, the application of the automated feature selection method was not needed for this study due to a limited number of independent variables, which were created from the original features based on their aggregation.

In addition to the number of features, another reason for this decision was that the points had the same impact on the final grade. The assignments and tests were evenly distributed in the first eight sections of the course, and their sequence did not play a significant role in this research.

The final reason is that, as was mentioned earlier, the course was used over four academic years. During this period, not all assignments and tests were always used. As a result, several features should be removed or require applying suitable imputation methods to be correctly included in the dataset. These approaches would result in a further reduction of an already limited dataset. Therefore, the partial results of different activities

were aggregated, reflecting that the partial assignments had the same share of the final grade in the courses in all academic years.

Simultaneously, alternative techniques of input feature aggregating based on course topics and weeks during the semester were examined. However, they were finally considered less perspective than the aggregation based on the categories.

As a result, the following five attributes reflecting the categories of the course activities were considered in the presented research (Table 1):

- *Accesses* represent the total number of course views by a student in the observed period.
- *Assignments* represent a total score from different types of evaluated activities within the observed period.
- *Tests* represent a total score from the midterm and final tests during the semester.
- *Exam* and *Project* features were also considered. While offering high analytic value, these were not included since they were realized in the last third of the term, and their values are directly tied to completing the course. Those values were not available at the time of expected prediction (during a semester when features *Project* and *Exam* are unknown yet).

**Table 1.** Descriptive statistics of independent variables for the dataset from years 2016–2019.

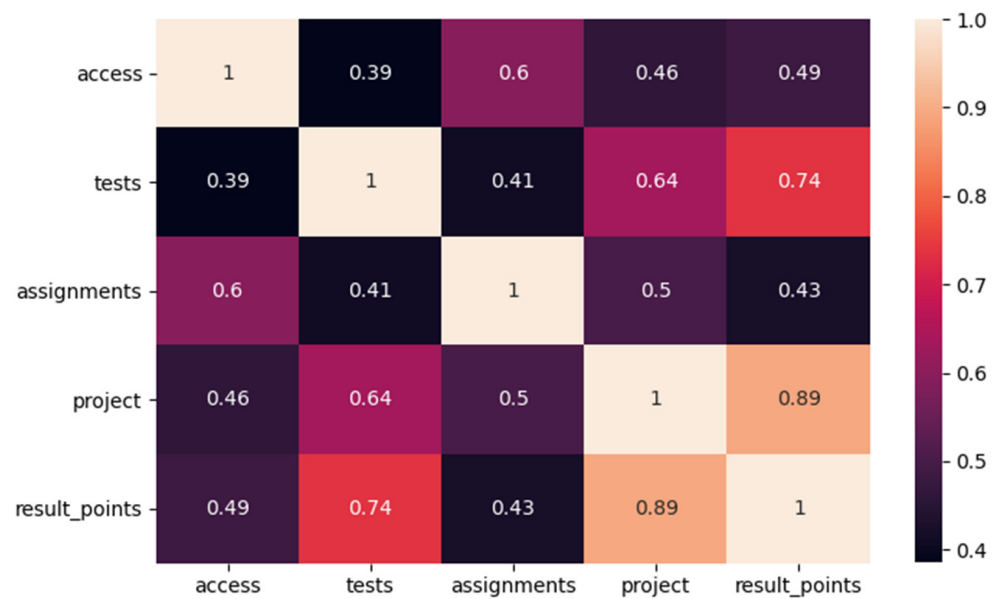
	Access	Tests	Assignments	Exam	Project
count	261.000	261.000	216.000	261.000	261.000
mean	680.295	37.941	18.077	11.555	65.513
std	374.662	13.113	10.831	6.303	30.615
min	13.000	0.000	0.000	0.000	0.000
25%	440.000	34.079	11.000	9.000	66.410
50%	605.000	40.500	14.400	14.000	75.140
70%	898.750	46.500	24.000	16.000	84.610
max	2392.000	57.000	40.000	20.000	99.480

A standard Pearson correlation was performed to find the correlations between the individual independent variables and final results, represented by the variable *result\_points*. This variable represents the total sum of partial points, which the student could get during the course. A correlation matrix, created using the Seaborn library in Python, is presented in Figure 2. The following conclusions can be drawn from the correlation values depicted in Figure 2.

- there is a medium correlation between the key attributes, with the values correlating to the resulting number of points at the end of the course. This finding is not surprising,
- variables *project* and *tests* have a slightly stronger correlation with *result\_points* than the other attributes. It can be deduced that they have a more critical impact on the final student's outcome. Therefore, they should be considered thoroughly concerning their occurrence in the course. If used later in the course, they would have a more significant impact on prediction accuracy.

Because the prediction should be made as early as possible to have enough time for intervention, the variables *access*, *assignments*, and *tests* were finally selected as the two most important features. It is necessary to emphasize, that only sum of partial tests, which were realized in the observed period, was taken into account. It can also be stated that a more appropriate attribute for prediction, besides the two mentioned attributes with a high correlation (*project* and *tests*), is also the number of accesses (*access*). One possible use of this attribute may be to identify students who do not achieve the minimum time allowance (presence or frequency of accesses), thus exposing them to a greater risk of obtaining a weaker point evaluation.





**Figure 2.** A correlation matrix between independent input variables.

A supervised machine learning model also requires an example of outputs (target) from which it can learn in addition to the input data. Since the information about the final grades of students was available, it was possible to create a binary target variable. Each student whose final grade was “FX” (failed) was assigned a numeric value of 0, while successful students were assigned a numeric value of 1. This way, the binary classification was achieved.

Data used in the case study consisted of two parts. The first part, used as a training and testing dataset, contained data from the three academic years (2016–2019). The second part was formed from the most recent run of the course in the year 2020. This data was used for evaluating the performance of the final classification models (Table 2).

**Table 2.** Descriptive statistics of independent variables for the dataset from the year 2020.

	Access	Tests	Assignments	Exam	Project
count	60.000	60.000	60.000	60.000	60.000
mean	925.533	45.114	25.533	11.842	74.405
std	309.606	14.320	10.357	5.192	33.238
min	260.000	0.000	0.000	0.000	0.000
25%	694.250	35.00	20.000	11.334	76.800
50%	846.000	49.525	28.000	13.375	88.470
70%	1152.500	56.360	34.00	14.901	94.555
max	1886.000	68.100	39.000	19.400	105.000

The evaluation dataset (Table 2) showed a substantial increase in the mean values overall compared to the data from Table 1. These changes in an absolute number of points were established in the last run of the course with the aim to practice the students’ skills. The mean values of the category project and tests increased by 14% and 18%, respectively. The categories access and assignments achieved considerable higher mean values of 36% and 42%. They had to be standardized before their use for evaluation of the proposed classification models.

Due to this fact, the final step of data preparation was normalisation and standardization of a dataset, which represent common requirements of many machine learning estimators. If the specific feature does not look like the standard data normally distributed, the estimators may behave incorrectly [33]. The data was transformed using Standard Scaler to standardize features by removing the mean and scaling to unit variance. Centring and scaling occurred independently on each feature by computing the relevant statistics on the dataset.

#### 4. Model

The modelling phase consists of applying different machine learning techniques to the dataset. The goal of the prediction is to create a model based on the students' current activities and accomplishments that attempts to predict learner failure and future performance. It is a typical classification problem, which a binary classification model or a binary classifier can solve in order to predict whether a student can complete the course or not.

Nowadays, there are many different machine learning classification methods, which can be used in order to find a suitable predictive model [3]. Considering the requirements resulting from the main aim of the paper and related works [28], the following most adopted classifiers were finally applied:

1. Logistic regression (LR) is a classification algorithm also known as logit regression, which can calculate a predicted probability for a dichotomous dependent variable based on one or more independent variables [7,34].
2. A decision tree (DT) is defined by recursively partitioning the input space and defining a local model in each resulting region of input space. This algorithm divides this complex classification task into a set of simple classification tasks [31].
3. Naïve Bayes classifier (NB) is a simple probabilistic classifier with strict feature independence assumptions based on Bayes' theorem [4,35].
4. Support Vector Machine (SVM) was initially designed for binary classification by constructing or setting an ideal hyperplane that divides the dataset into classes [34].
5. Random Forest (RF) is a technique that uses multiple decision trees designed to perform the estimate using bootstrapping, resampling, and then applying majority voting or averaging [2,31].
6. A neural network (NN) is a parallel processing system that takes advantage of the brain's structure and functions to maximize predictive accuracy. Deep learning NN takes data as input and then processes the data to compute the output across various layers [31,33].

The main idea is to find the most suitable prediction model and compare the various prediction performance metrics of each applied classifiers. Hyperparameters were tuned using the grid search technique, which is widely used for finding out the best settings of parameters.

Python programming language was used. Python offers rich and efficient libraries for applying machine learning methods like Pandas and Numpy libraries for data manipulation. Subsequently, the Scikit-learn library containing various set of supervised or unsupervised machine learning methods was used. Other libraries, Keras and Tensorflow, were used for constructing a neural network. Keras is an API for Tensorflow, making it easier to build NNs [32,36].

Nowadays, there are available several low-code machine learning libraries that automate machine learning workflow. Their application in LA domain is not frequent. Therefore, considering the main aim of the article, which discuss the limitations of the application of the classifiers on the scarce educational dataset, several classifiers added using the PyCaret library [37].

It is challenging to create a suitable precise prediction point to identify students at risk of dropping out of a course if the dataset is smaller. The classification methods mentioned above were compared using data divided into training and testing sets. Testing and training sets were chosen randomly in order to obtain confidence intervals for all the following performance metrics [34]:

*Accuracy* is the overall accuracy rate or classification accuracy, and it is determined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

*True positive rate (TP rate) or recall* is the proportion of real positives predicted to be positive. In this case study, the *recall* was used to evaluate the completers. It is determined as follows:

$$Recall = \frac{TP}{TP + FN}$$

*True negative rate (TN rate) or precision* is the proportion of real negatives expected to be negative. In this paper, *precision* was used to assess the dropout of students, and it is determined as follows:

$$Precision = \frac{TN}{TN + FP}$$

$F1_{score}$  shows the balance between two measures of classification. It represents a measure widely used trade-off calculation for imbalanced data sets, and it is determined as follows:

$$F1_{score} = \sqrt{Recall * Precision}$$

Finally, the *classification error rate* represents a proportion of instances misclassified over the whole set of instances. It is calculated as follows:

$$e = \frac{FP + FN}{TP + TN + FP + FN}$$

This study should consider several metrics to estimate the performance of the different methods to correctly classify the cases in an independent dataset and avoid overfitting.

All classification algorithms were executed using a 10-fold cross-validation procedure. The 10-fold cross-validation process divides the collection of data into ten approximately equal sections. The nine remaining parts are used to train the model, and the test error is computed by classifying the given part. Lastly, the results of ten tests are averaged.

## 5. Results and Evaluation

The models produced in the previous phase should be evaluated for satisfactory implementation at this stage of CRISP-DM methodology. The evaluation used in this case study was numerical based on model performance metrics, in contrast to empirical evaluation, in which the creator of the model and the domain knowledge expert assess if the model fits the purpose.

The most recent data of the course from the academic year 2020 were used for evaluating the performance of each classification algorithms. This data is later referred to as unseen data.

### 5.1. Performance Metrics Comparison

Table 3 compares the 10-fold cross-validation test classification results obtained by Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Support Vector Machines (SVM), and Neural Network (NN) algorithms for the selected course, reporting on some of the most popular measures/indicators of success: accuracy, precision, recall, and the F1 score.

The Naïve Bayes classifier obtained the highest precision results on the evaluation dataset with a 93% success rate. However, the classification error was 0.231, with the overall accuracy of the model 77%. Considering this fact, this algorithm may have suffered from the strong independence assumption on the dataset and identified many false negative cases as the recall value reached only 72%, which reduced the overall accuracy. The F1 score reached 82%. Therefore, it is not suggested as a suitable classifier for predicting student's dropout in this case.

**Table 3.** Comparison of the performance measures of applied classification algorithms.

Model	Accuracy	Recall	Precision	F1 Score
Naïve Bayes (NB)	0.77	0.72	0.93	0.82
Random Forest (RF)	0.93	0.96	0.86	0.91
Neural Network (NN)	0.88	0.89	0.86	0.88
Logistic Regression (LR)	0.93	0.98	0.79	0.90
Support Vector Machines (SVM)	0.92	0.96	0.79	0.88
Decision Tree (DT)	0.90	0.98	0.71	0.85

According to Table 3, the random forest classifier obtained the best accuracy, precision, F1 scores and the second-best recall value. It is worth to mention that RF gains an advantage over other classifiers thanks to various hyperparameters that can be adjusted and lead to more accurate prediction. The resulting hyperparameters from the grid search pointed out the construction of 1000 number of trees within the forest model with a maximum depth of 50 for each tree. Thanks to these settings, the precision of the model reached 86%, the recall was 96%, while the F1 score was 91%. The classification error was 0.066, with an overall accuracy of 93%.

The neural network model constructed in Keras is described as a series of layers. The ReLU activation function was used in both hidden layers. Due to the binary classification problem, the sigmoid activation function was used in the output layer to ensure output values between 0 and 1. Optimizer “adam” has been chosen since it automatically tuned itself and gave good results in a wide range of problems. Considering that the dataset is relatively small, and the training process will go through for a fixed number of iterations, the number 50 was picked with a batch size of 10, which is the number of dataset rows considered before the model weights are changed for this problem. With all those settings, the classification error was 0.122, with an overall accuracy of 88%. The neural network achieved an F1 score of 88%. Recall reached 89%, while precision ended with 86% accuracy.

There are no critical hyperparameters to tune within the logistic regression classifier. No useful differences in performance were achieved with various solvers, regularization, and C parameter, which controls the regularization strength. The overall accuracy increased by 2–3%, but the recall value significantly decreased during hyperparameters optimization. Therefore, in this case, the final results achieved by LR were obtained under the default settings where the scores for precision and recall were 79% and 98%, respectively. The classification error was 0.067, with an overall accuracy of 93%.

The support vector machine classification algorithm is very common and has several hyperparameters to tune up the outcome. The choice of kernel, which will control how the input variables are projected, is perhaps the most important parameter. The C parameter should also be considered as important to choose. It represents the penalty and can contain a variety of values. It has a significant impact on the shape of the resulting regions for each class. In this case, the resulting hyperparameters obtained by grid search pointed out to linear kernel with 1.0 penalty value. The classification error was 0.073, with an overall accuracy of 92%. The algorithm achieved an F1 score of 88%, while recall and precision scores reached 96% and 79%, respectively.

The tree size, also known as the number of layers, is the most critical decision tree model parameter. Shallow trees, also known as weak learners, are expected to perform poorly because they do not capture enough information about the problem. On the other hand, deeper trees capture too many aspects, which leads to overfitting the training dataset, making it difficult to make an accurate prediction on unseen data. As a result, the decision tree model was created under the depth parameter of 4. This model obtained 90% overall accuracy and classification error was 0.103. While recall reached a promising value of 98%, the precision achieved only 71%, which also affected the lower F1 score of 85%.

### 5.2. Comparison of Used Machine Learning Classifiers

The performance of the selected classifiers can also be compared using statistical significance tests for comparing machine learning algorithms, even though the dataset used in this case study was small and training multiple copies of classifier models were not expensive or impractical. As each model was evaluated on the same dataset, McNemar's test could be used to compare the classification algorithms if the models seem to predict the same way or not. McNemar's test is a paired nonparametric or distribution-free statistical hypothesis test [38]. This test checks if the disagreements between two cases match. Therefore, McNemar's test is a type of homogeneity test for contingency tables.

The results of McNemar's test (Table 4) show that paired classifiers with a  $p$ -value higher than 0.5 had consistently underpredicted and overpredicted the same test set. It means the classifiers have a similar error proportion on the test set. Simultaneously, the results show that they are not significant in the majority of cases.

**Table 4.** McNemar's test comparison of the classifiers.

Model	LR	DT	RF	NB	SVM
LR	-	$p = 1.000$	$p = 0.500$	$p = 0.001$	$p = 1.000$
DT	-	-	$p = 0.250$	$p = 0.000$	$p = 1.000$
RF	-	-	-	$p = 0.007$	$p = 0.250$
NB	-	-	-	-	$p = 0.000$
SVM	-	-	-	-	-

On the other hand, all classifiers paired with the NB model had obtained a  $p$ -value lower than 0.5. If different models make different errors, one model is more accurate than the other and most likely has a different proportion of errors on the test set. In that case, the result of the test is significant. Therefore, as it was already assumed, the NB classifier is considered not suitable for predicting student's dropout in this case.

## 6. Discussion

Six of the well-known machine learning algorithms were compared using a dataset about the activity of 261 third-year students in the university course. The dataset was pre-processed in a standard way. The models were trained and tested on the dataset from three academic years. The findings obtained are based on a balanced dataset. Measures of success have been recorded in both categories, completers and non-completers. Both biases were avoided in terms of an unbalanced dataset as well as based on averaging. Subsequently, the models were evaluated unseen data from the fourth academic year. The classifiers were evaluated using several performance metrics and using McNemar's test, which compared their ability to classify the unseen cases in the same classes.

The best dropout prediction based on the comparison of the precision, recall, accuracy and F1 scores was obtained by random forest (RF) classifier in general. The results obtained by the LR, SVM, and DT algorithms were the second-best with slight differences compared to RF.

On the other hand, the Naïve Bayes and the Neural Network model had the worst results overall and are not considered accurate classification models for this limited dataset. This finding confirms that the performance of the supervised machine learning models closely depends on enough records. The evaluation data was slightly different from the training data because the course has received minor changes compared to the previous runs. The NN's accuracy was mostly affected by these changes as the dataset consists of relatively few records. Standardization or normalization of the dataset was considered to improve the NN model. Finally, standardization of attributes was chosen as it is recommended and useful for models that depend on attribute distribution.

The prediction accuracy varied between 77 and 93%, indicating that regardless of the model used, the chosen features in this study proved to be successful in predicting student's success or dropout even though the limitation of the dataset and number of features.

On the other hand, all algorithms achieved recall values higher than 0.77 and simultaneously a very good precision. However, classification algorithms could achieve very high accuracy by only predicting the majority class. For instance, if any of the used algorithms predict all students as “completers ‘1’”, the accuracy of the dataset is expected to be 78.5% because 21.5% of students dropped out, caused by the statistical distribution of the data.

The examined predictive models were capable of reaching high performance in each class (completers ‘1’ and non-completers ‘0’) considering the values of recall and precision, as shown in Table 3. The current accuracy of the classification algorithms’ predictions at this stage should be trusted.

To sum up, there is a convergence between the metrics of RF, LR, SVM, and DT models. In addition, the harmonic mean of precision and recall was represented by the F-score. In addition, the F-score represented the harmonic mean of precision and recall, which indicates the overall success rate of prediction models to predict students who failed. Considering the F-score values shown in Table 3, the RF, LR, NN, and SVM models were comparatively useful in predicting students’ likelihood to drop out.

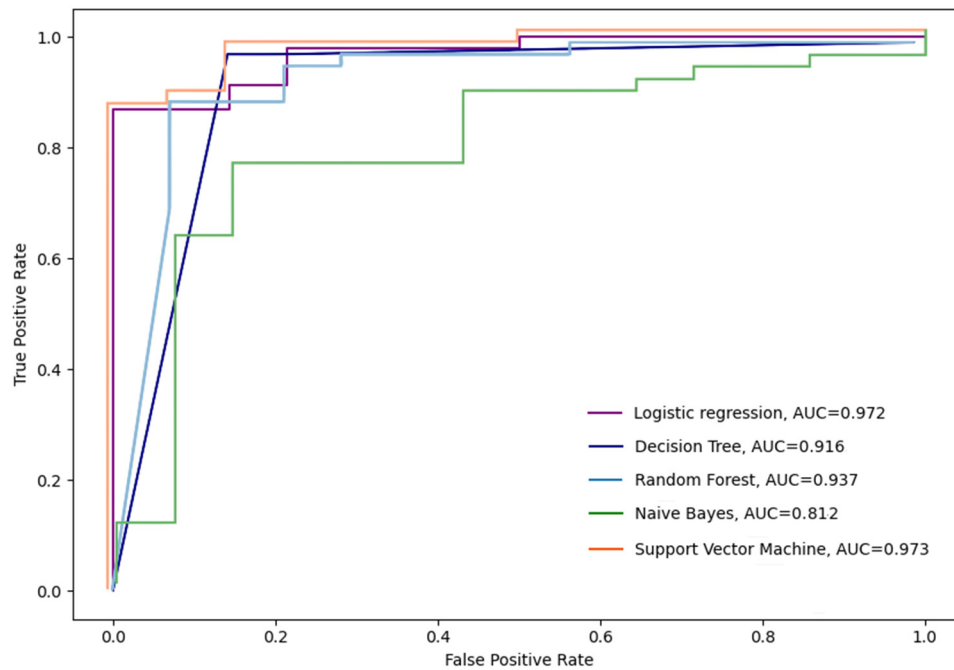
The performance of proposed classification algorithms was measured for further analysis regarding the effectiveness of various proposed models. The area under the receiver operating characteristic curve (AUC-ROC) was considered to evaluate the impact of an imbalanced dataset. The AUC-ROC is typically used to address imbalanced data because it demonstrates the trade-off between the true positive rate (TPR) and the false positive rate (FPR) in a two-dimensional space [23].

The ROC curve is obtained by plotting the values of the various thresholds calculated. The TPR represents the percentage of students’ dropout that the model predicted as dropout states in this research. Vice versa, the FPR is the percentage of students’ non-dropouts, but the classifier predicted them as a dropout. This metric summarizes the goodness of fit of a classifier [30].

Figure 3 shows that random forest (RF) (AUC = 0.968) and logistic regression (LR) (AUC = 0.924) classifiers are significantly better when measured by the AUC-ROC curve. Decision tree (DT) gained (AUC = 0.883), followed by a support vector machine (SVM) (AUC = 0.850) and Naïve Bayes (NB) (AUC = 0.765).

Considering the distribution of positive class (completers ‘1’) and negative class (non-completers ‘0’) in the researched dataset, each of the used classification algorithms is very good at capturing positive class occurrences. However, some of them, such as DT, SVM, and NB, generate lots of false-positive predictions, which lowers final predictive performance accuracy. Random forest classifier and logistic regression can distinguish little differences between the conditioning parameters that result in fewer incorrectly predicted failures. Therefore, their predictive performance is higher.

This statement can also confirm the classification error rate, representing a proportion of cases misclassified over the whole set of cases. Again, the RF and LR classifiers achieved the best values in this performance metric. They misclassified only around 6% of the cases, which is in stark contrast to the NB classifier, which misclassified 23% of the cases.



**Figure 3.** Comparison of the performance measures by area under the receiver operating characteristic curve (AUC-ROC) curves.

McNemar’s test provides another interesting point of view on the individual classifiers. The mutual comparison of the binary classifiers shown in Table 4 uncovered that the RF and LR classifiers classify the individual unseen cases to the same classes. There is no significant difference between them. Simultaneously, there is no significant difference between other classifiers, namely, between DT and SVM, DT, and LR, and between LR and SVM.

Finally, the computational cost of running all the research algorithms was compared for completeness sake. The four classification algorithms were executed in less than 1.00 s, as shown in Table 5. Naturally, the NN recorded a significantly greater time because of hidden layers. However, given the time frame of the proposed problem, its runtime was not prohibitive. It took around nine seconds to run while using all the attributes in the worst case.

**Table 5.** Execution time (in seconds).

Model	Execution Time
LR	0.36
DR	0.84
RF	0.46
NB	0.38
SVM	0.52
NN	8.95

The discussion about the different performance metrics confirms that the selected binary classification models can be used for predicting students’ success or dropout at the individual course level, even in the case of the scarce dataset with a limited number of input features. However, it is necessary to explore a more extensive set of performance metrics before the classifiers should be deployed for students’ dropout prediction. This statement is in line with the outcomes of other research papers published in the learning analytics and educational data mining domain.

Nevertheless, the presented case study and its findings have some limitation. As was already mentioned, the limited size of the dataset is the first one. In contrast to many other application domains of machine learning models, the amount of data in the educational domain cannot be easily increased by combining different resources. The

reason is that individual records should often represent the learning outcomes or behavior of individual students. As a consequence, the datasets are often smaller than the machine learning algorithms would really need. This shortcoming can be often overcome only by applying the similar precise research design used in the “classical” educational technology research, in which a sufficient number of records is estimated or guaranteed before the experiment itself.

The same requirement is also true for the number and the quality of the independent variables (features). The learning process often relies on the repetition of some activities. If these activities are mutually connected or conditioned, their inclusion in the machine learning technique should be thoroughly scoped, again, before the process of data collection starts.

The selection of the time threshold to which the prediction is made can also be mentioned as a weakness. Since the time variable was not directly included, it was necessary to find milestones when the students’ performance would not be easily predictable. The end of the second third of the term was used as a compromise concerning the natural distribution of the individual categories of activities in the course sections. Nevertheless, better results would be expected if the activities in the course sections were designed knowing the requirements of the applied machine learning techniques.

Another limitation of this research is that individual runs of the courses offered different data that could be analyzed. Therefore, it was challenging to figure out which attributes are relevant enough for the student’s performance prediction in general. For this purpose, the CRISP-DM methodology was used to repeat the individual research steps and discover hidden patterns and meaningful relationships within a selected dataset or understand the data that has the most significant impact on solving the dropout problem. Simultaneously, the maximum range of the points, which could the students achieve, should be analyzed and standardized.

The selection of the classifiers used in this case study can be considered the next limitation. Literature reviews of the current state of the research and trends in LA and EDM provide many examples of more or less advanced machine learning techniques, which can be applied for early students’ dropout prediction at the course level. However, any of them have not provided significantly better results so far. Since the main aim of the paper was not to find the best one, the final selection of the classifier used in this case study allows pointing out that good prediction could also be achieved using simpler classifiers, but the performance metrics must be evaluated.

The last weakness of the case study is a form of suitable intervention, which is not discussed in detail. The case study has already shown that the student’s failure can be predicted based on the selected categories of activities. On the other hand, the reason for this unflattering state remains open and requires further research. It would be interesting to examine whether other categories of activities should have a similar impact on students’ engagement and for which categories of activities they can be exchanged in the intervention phase.

## 7. Conclusions

The students’ dropout rate is one of the primary performance indicators of e-learning. It is crucial in the case of the virtual learning environment at the university, as well as in the case of MOOCs (*Massive Open Online Course*). The reliable prediction can encourage students to study harder when they know that they are at risk and plan their semester workload carefully. The school management can take this information into account when deciding whether to confirm or reject students’ requests who appeal for repeating failed online courses. Predicting students at-risk of dropout also gives educators a signal to intervene when students are at risk and improve students’ involvement in interactive activities in time. It is essential to reveal what courses have in common and how this set of features can influence students’ behavior and performance.



The presented research contributes to solving the problem of students' dropouts at the course level. The results showed that suitably selected indicators, which do not require access to the system logs, may be beneficial despite a small dataset if different performance metrics are evaluated. Gathered data about students' online learning environment activities and their partial achievements were used to input the predictive models. Simultaneously, a proposed methodology has been shown to be reliable for predicting course completion when there is enough time to encourage educators to make timely interventions.

The results discovered that the classification models could be considered trustworthy enough to make a proper prediction of completing or early dropout the course before the end of a semester. In fact, good results of prediction were obtained, regardless of the employed model. The features selected in this study proved to be influential in the prediction of completer and non-completers. The prediction accuracy varied between 77 and 93%.

In this context, it is important to note that recognizing students at risk of dropping out is just the first step in addressing school dropout. The next step is to recognize each student's unique needs and problems dropping out and then introduce adjustments to include effective dropout prevention strategies [39]. This can illustrate the real reasons for students' dropout and develop strategies that may encourage students to complete their courses on time successfully. Instructors should also be able to focus on students' needs to help them prevent dropouts in time.

At the same time, it is also worth remembering that the secret to assessing the impact of these various types of interventions is to find the most suitable one for each form of dropout-risk student. Therefore, future research should focus on applying proposed models to predict students at risk of dropping out based on similar students' previous achievements. The knowledge obtained from this research may serve as the basis for designing a course recommendation.

These guidelines, however, should take into account the talents, skills, preferences, and free-time activities of students in their schedules. The proposed methods with easily obtainable predictors are rather general and can be utilized on other courses.

**Author Contributions:** Conceptualization, J.K. and M.D.; methodology, J.K. and M.D.; validation, J.K. and M.D.; formal analysis, J.K. and M.D.; resources, J.K. and M.D.; data curation, J.K. and M.D.; writing—original draft preparation, J.K. and M.D.; writing—review and editing, J.K. and M.D.; visualization, Janka Kabathova; supervision, M.D.; project administration, M.D.; funding acquisition, M.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the European Commission under the ERASMUS+ Programme 2018, KA2, grant number: 2018-1-SK01-KA203-046382 "Work-Based Learning in Future IT Professionals Education" and the Cultural and Educational Agency of the Ministry of Education of the Slovak Republic, grant number: KEGA029UKF-4/2018 "Innovative Methods in Programming Education in the University Education of Teachers and IT Professionals".

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The data presented in this study are openly available in <https://github.com/JKabathova/Students-Dropout-Prediction> (accessed on 10 February 2021).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Narayanasamy, S.K.; Elçi, A. An Effective Prediction Model for Online Course Dropout Rate. *Int. J. Distance Educ. Technol.* **2020**, *18*, 94–110. [CrossRef]
2. Wang, W.; Yu, H.; Miao, C. Deep Model for Dropout Prediction in MOOCs. In Proceedings of the 2nd International Conference on Cryptography, Security and Privacy, Beijing, China, 6–9 July 2017; pp. 26–32. [CrossRef]
3. Prenkaj, B.; Velardi, P.; Stilo, G.; Distanto, D.; Faralli, S. A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses. *ACM Comput. Surv.* **2020**, *53*, 1–34. [CrossRef]

4. Queiroga, E.M.; Lopes, J.L.; Kappel, K.; Aguiar, M.; Araújo, R.M.; Munoz, R.; Villarroel, R.; Cechinel, C. A Learning Analytics Approach to Identify Students at Risk of Dropout: A Case Study with a Technical Distance Education Course. *Appl. Sci.* **2020**, *10*, 3998. [[CrossRef](#)]
5. Lu, X.; Wang, S.; Huang, J.; Chen, W.; Yan, Z. What Decides the Dropout in MOOCs? In Proceedings of the International Conference on Database Systems for Advanced Applications, Suzhou, China, 27–30 March 2017; pp. 316–327. [[CrossRef](#)]
6. Yang, B.; Qu, Z. Feature Extraction and Learning Effect Analysis for MOOCs Users Based on Data Mining. *Educ. Sci. Theory Pract.* **2018**, *18*, 1138–1149. [[CrossRef](#)]
7. Moreno-Marcos, P.M.; Alario-Hoyos, C.; Munoz-Merino, P.J.; Kloos, C.D. Prediction in MOOCs: A Review and Future Research Directions. *IEEE Trans. Learn. Technol.* **2019**, *12*, 384–401. [[CrossRef](#)]
8. Mubarak, A.A.; Cao, H.; Zhang, W. Prediction of students' early dropout based on their interaction logs in online learning environment. *Interact. Learn. Environ.* **2020**. [[CrossRef](#)]
9. Jin, C. MOOC student dropout prediction model based on learning behavior features and parameter optimization. *Interact. Learn. Environ.* **2020**. [[CrossRef](#)]
10. Drlik, M.; Munk, M.; Skalka, J. Identification of Changes in VLE Stakeholders' Behavior Over Time Using Frequent Patterns Mining. *IEEE Access* **2021**, *9*, 23795–23813. [[CrossRef](#)]
11. Shaun, R.; Baker, J.D.; Inventado, P.S. *Educational Data Mining and Learning Analytics*; Springer: Berlin/Heidelberg, Germany, 2014; Chapter 4; pp. 61–75.
12. Siemens, G.; Baker, R.S.J.D. Learning analytics and educational data mining. In Proceedings of the 2nd International Conference on Cryptography, Security and Privacy, Vancouver, BC, Canada, 29 April–2 May 2012; pp. 252–254.
13. Alamri, A.; AlShehri, M.; Cristea, A.; Pereira, F.D.; Oliveira, E.; Shi, L.; Stewart, C. Predicting MOOCs Dropout Using Only Two Easily Obtainable Features from the First Week's Activities. In Proceedings of the International Conference on Intelligent Tutoring Systems, Kingston, Jamaica, 3–7 June 2019; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2019; Volume 11528, pp. 163–173.
14. Romero, C.; Ventura, S.; Baker, R.; Pechenizkiy, M. *Handbook of Educational Data Mining (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*; CRC Press: Boca Raton, FL, USA, 2010; ISBN 1584888784.
15. Skalka, J.; Drlik, M. Automated Assessment and Microlearning Units as Predictors of At-risk Students and Students' Outcomes in the Introductory Programming Courses. *Appl. Sci.* **2020**, *10*, 4566. [[CrossRef](#)]
16. Ifenthaler, D.; Yau, J.Y.-K. Utilising learning analytics to support study success in higher education: A systematic review. *Educ. Technol. Res. Dev.* **2020**, *68*, 1961–1990. [[CrossRef](#)]
17. Drlik, M.; Munk, M. Understanding Time-Based Trends in Stakeholders' Choice of Learning Activity Type Using Predictive Models. *IEEE Access* **2018**, *7*, 3106–3121. [[CrossRef](#)]
18. Hämäläinen, W.; Vinni, M. Classifiers for educational data mining. In *Educational Data Mining Handbook*; Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.D., Eds.; CRC Press: Boca Raton, FL, USA, 2010; pp. 57–74.
19. Lang, C.; Siemens, G.; Wise, A.; Gasevic, D. (Eds.) *Handbook of Learning Analytics*; SOLAR: Beaumont, AB, Canada, 2017.
20. Kloft, M.; Stiehler, F.; Zheng, Z.; Pinkwart, N. Predicting MOOC Dropout over Weeks Using Machine Learning Methods. In Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, Doha, Qatar, 25–29 October 2014; pp. 60–65. [[CrossRef](#)]
21. Uden, L. Learning Technology for Education in Cloud. *Commun. Comput. Inf. Sci.* **2015**, *533*, 43–53. [[CrossRef](#)]
22. Baneres, D.; Rodriguez-Gonzalez, M.E.; Serra, M. An Early Feedback Prediction System for Learners At-Risk Within a First-Year Higher Education Course. *IEEE Trans. Learn. Technol.* **2019**, *12*, 249–263. [[CrossRef](#)]
23. Kennedy, G.; Coffrin, C.; De Barba, P.; Corrin, L. Predicting success. In Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction, Poughkeepsie, NY, USA, 16–20 March 2015; pp. 136–140. [[CrossRef](#)]
24. Olivé, D.M.; Huynh, D.Q.; Reynolds, M.; Dougiamas, M.; Wiese, D. A supervised learning framework: Using assessment to identify students at risk of dropping out of a MOOC. *J. Comput. High. Educ.* **2019**, *32*, 9–26. [[CrossRef](#)]
25. Benko, L.; Reichel, J.; Munk, M. Analysis of student behavior in virtual learning environment depending on student assessments. In Proceedings of the 2015 13th International Conference on Emerging eLearning Technologies and Applications (ICETA), Stary Smokovec, Slovakia, 26–27 November 2015; IEEE: Piscataway, NJ, USA; pp. 1–6.
26. Herodotou, C.; Rienties, B.; Boroowa, A.; Zdrahal, Z.; Hlosta, M.; Naydenova, G. Implementing predictive learning analytics on a large scale. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, BC, Canada, 13–17 March 2017; pp. 267–271. [[CrossRef](#)]
27. Márquez-Vera, C.; Cano, A.; Romero, C.; Noaman, A.Y.M.; Fardoun, H.M.; Ventura, S. Early dropout prediction using data mining: A case study with high school students. *Expert Syst.* **2016**, *33*, 107–124. [[CrossRef](#)]
28. Charitopoulos, A.; Rangoussi, M.; Koulouriotis, D. On the Use of Soft Computing Methods in Educational Data Mining and Learning Analytics Research: A Review of Years 2010–2018. *Int. J. Artif. Intell. Educ.* **2020**, *30*, 371–430. [[CrossRef](#)]
29. Romero, C.; Espejo, P.G.; Zafra, A.; Romero, J.R.; Ventura, S. Web usage mining for predicting final marks of students that use Moodle courses. *Comput. Appl. Eng. Educ.* **2010**, *21*, 135–146. [[CrossRef](#)]
30. Rastrollo-Guerrero, J.L.; Gómez-Pulido, J.A.; Durán-Domínguez, A. Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Appl. Sci.* **2020**, *10*, 1042. [[CrossRef](#)]

31. Xing, W.; Chen, X.; Stein, J.; Marcinkowski, M. Erratum: Corrigendum to “Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization” (Computers in Human Behavior (2016) 58 (119–129)(S074756321530279X)(10.1016/j.chb.2015.12.007)). *Comput. Human Behav.* **2017**, *66*, 409. [[CrossRef](#)]
32. Youssef, M.; Mohammed, S.; Hamada, E.K.; Wafaa, B.F. A predictive approach based on efficient feature selection and learning algorithms’ competition: Case of learners’ dropout in MOOCs. *Educ. Inf. Technol.* **2019**, *24*, 3591–3618. [[CrossRef](#)]
33. Obonya, J.; Kapusta, J. Identification of Important Activities for Teaching Programming Languages by Decision Trees. In Proceedings of the 12th International Scientific Conference on Distance Learning in Applied Informatics (DIVAI), Štúrovo, Slovakia, 2–4 May 2018; Turcani, M., Balogh, Z., Munk, M., Kapusta, J., Benko, L., Eds.; Kluiwert: Sturovo, Slovakia, 2018; pp. 481–490.
34. Hagedoorn, T.R.; Spanakis, G. Massive Open Online Courses Temporal Profiling for Dropout Prediction. In Proceedings of the 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), Boston, MA, USA, 6–8 November 2017; pp. 231–238.
35. Lacave, C.; Molina, A.I.; Cruz-Lemus, J.A. Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks. *Behav. Inf. Technol.* **2018**, *37*, 993–1007. [[CrossRef](#)]
36. Doleck, T.; Lemay, D.J.; Basnet, R.B.; Bazelais, P. Predictive analytics in education: A comparison of deep learning frameworks. *Educ. Inf. Technol.* **2020**, *25*, 1951–1963. [[CrossRef](#)]
37. Ali, M. PyCaret: An Open Source, Low-Code Machine Learning Library in Python. 2021. Available online: <https://pycaret.readthedocs.io/en/latest/index.html> (accessed on 10 February 2021).
38. Dietterich, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [[CrossRef](#)] [[PubMed](#)]
39. Ferguson, R.; Cooper, A.; Drachler, H.; Kismihók, G.; Boyer, A.; Tammets, K.; Monés, A.M. Learning analytics. In Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction, Poughkeepsie, NY, USA, 16–20 March 2015; pp. 69–72.