




Article

A Case Study on the Data Mining-Based Prediction of Students' Performance for Effective and Sustainable E-Learning

Evelina Staneviciene ¹, Daina Gudoniene ^{1,*}, Vytenis Punys ¹ and Arturas Kukstys ²

¹ Faculty of Informatics, Kaunas University of Technology, Studentu Str. 50, LT-44249 Kaunas, Lithuania; evelina.staneviciene@ktu.lt (E.S.); vytenis.punys@ktu.lt (V.P.)

² Bergen Kommune, Rådhusgaten 10, Bergen Rådhus, 5014 Bergen, Norway; arthuras.kukstys@bergen.kommune.no

* Correspondence: daina.gudoniene@ktu.lt

Abstract: The study explores the application of data analytics and machine learning to forecast academic outcomes, with the aim of ensuring effective and sustainable e-learning. Technological study programs in universities often experience high dropout rates, which makes it essential to analyze and predict potential risks to reduce dropout percentages. Student performance prediction (SPP) offers potential benefits, including personalized learning and early interventions. However, challenges such as (1) data quality and availability and (2) incomplete and inconsistent data complicate this process. Moreover, to support the fourth Sustainable Development Goal (SDG), we focus on the quality of education. A case study approach is used using data mining techniques, particularly classification, regression, and clustering, to predict student performance. The case presented aims to predict risks and ensure academic success and quality. The cross-industry standard process for data mining (CRISP-DM) methodology is used to structure and guide the prediction process. The study shows that using data from student learning processes within an academic success prediction model and data mining can identify at-risk students.

Keywords: educational data mining; prediction; academic success; student performance; sustainable development goals; machine learning



Citation: Staneviciene, E.; Gudoniene, D.; Punys, V.; Kukstys, A. A Case Study on the Data Mining-Based Prediction of Students' Performance for Effective and Sustainable E-Learning. *Sustainability* **2024**, *16*, 10442. <https://doi.org/10.3390/su162310442>

Academic Editors: Luis Ortiz Jiménez and Tai-Yi Yu

Received: 23 October 2024

Revised: 21 November 2024

Accepted: 26 November 2024

Published: 28 November 2024



Copyright: © 2024 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

This study presents a case study of data mining in an engineering study program as a modern data analysis process that provides the opportunity to extract useful information from accumulated data, making it suitable for the management of the analyzed activity, problem analysis, decision-making, prediction, etc. The emergence of data mining was driven by the imperfections within classical statistical methods and advances in artificial intelligence and machine learning. This method resembles statistics as statistics and data mining are both data analysis-oriented processes that require the organization of “raw” data, but it should also be noted that data mining should not be equated with statistics. Statistical analysis is generally applied to primary data analysis and data research as well as secondary data analysis [1]. According to Manjarres et al. [2], data mining can be viewed as a set of methods and procedures designed to analyze large amounts of data, such as transfer transactions, scientific research data, personal health data, videos and photos, data recorded by satellites, etc., stored in various databases.

The use of data mining techniques to help extract and analyze large amounts of data in educational sectors to improve teaching and learning processes is called educational data mining [3]. Educational data mining (EDM) is defined as the extraction of new information from large amounts of educational data collected in the educational environment and stored in educational databases [4]. EDM is an area of study that focuses on the use of techniques such as data mining, machine learning, and statistical analysis to extract meaningful

information from complex datasets [5]. EDM includes processes such as collecting data, applying models to describe those data, and obtaining useful information about students.

This study uses educational data mining as it is useful for understanding student learning behavior to develop teaching strategies that improve student performance and reduce dropout rates [3]. Another area closely related to educational data mining is learning analytics (LA).

Both EDM and LA are interdisciplinary fields that include data retrieval, visual data analysis, domain-based data mining, social sciences, psychology and cognitive science, etc. The authors [4] define these fields as a combination of computer science, education, and statistics (see Figure 1).

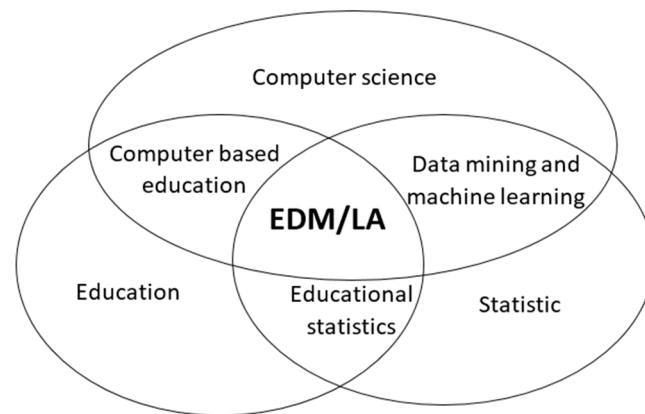


Figure 1. Areas related to educational data mining and learning analytics.

Figure 1 illustrates the interdisciplinary framework of educational data mining and learning analytics, highlighting the integration of education, information technology, statistics, and computer-based education. Education provides theories of learning and teaching, computer science offers technical methods for developing analytical tools, statistics allow you to analyze data and recognize patterns, and computer-based education focuses on the use of technology for learning. In the center, EDM/LA combines data mining and machine learning to improve learning outcomes through insights gained from these interrelated fields [4].

Different methods such as classification, regression, and clustering are generally used in educational data mining [6,7]. The method of association rules, the method of sequencing research, and the method of data visualization can also be applied, which allows the data to be displayed understandably and clearly (see Figure 2).

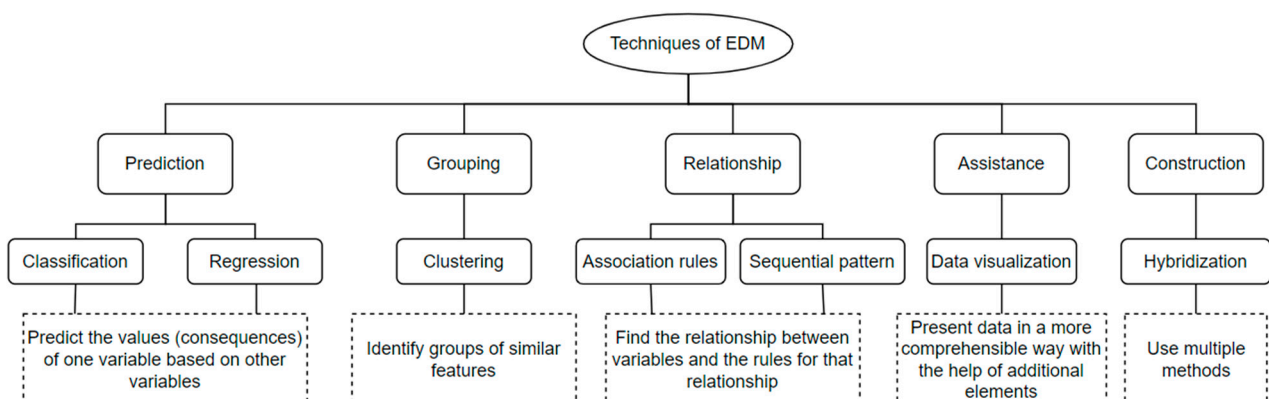


Figure 2. Educational data mining techniques.

Prediction techniques are used to predict the probability that learners will pass/fail an exam or complete/fail a module, course, or study. In this case, a classification method can be used [8]. By training the model on historical data, it learns patterns and relationships that allow it to make probabilistic predictions of learners' exam performance. A linear regression method is used to predict the academic performance of learners [9]. Clustering methods determine which learning materials should be improved and which learning materials learners should choose when preparing for exams [10]. By clustering data such as learners' performance on different materials or topics, it becomes possible to identify clusters where performance is consistently low, indicating areas where the materials may require enhancement to better support learning outcomes. J. Chen and J. Zhao [11] used data on the learning processes of learners and applied the association rule method to determine which learning habits help learners to learn English. Finding sequential patterns allows us to define patterns of learner behavior that lead to a particular learning outcome [12]. Data visualization can be used to show how quickly a certain learning material is learned and to help understand learner learning patterns, outcomes, etc. [13]. It is also possible to use multiple models, such as first applying clustering to a group of learners and then using classification to predict the achievement of an individual learner [14].

Student performance can be predicted through interactions with learners, surveys and assessments, and educational data mining [15,16]. In the literature [7,17], academic success is defined as a multidimensional concept that includes academic achievements, involvement in the learning process, satisfaction experienced during learning, acquired competencies and skills during learning, overcoming learning difficulties, continuing learning, favorable professional career development, and the achievement of learning goals. Communication or assessment activities within study programs can be organized and implemented in a virtual learning environment, such as Moodle, Google Classroom, or others. It should also be mentioned that students with a high academic self-efficacy score better and graduate successfully, so academic self-efficacy is considered one of the most important psychological characteristics for predicting academic success [18,19]. In other words, academic self-efficacy and learning achievements in academic activities are closely related [20].

Here, various learner types of data are analyzed (learner actions in the virtual learning environment, responses to psychological surveys, demographic characteristics, etc.) and information is sought about the risk of academic failure in predicting the academic success of learners. The authors of References [21,22] distinguish between two main types of data that are used to predict academic success: (1) administrative data and (2) learning process data. The most valuable information in educational data mining is obtained when the educational datasets under study contain both types of data.

This paper aims to provide a case for predicting learners' academic success by applying educational data mining methods to reduce student dropout in the future. The authors of References [23,24] recommend using the CRISP-DM data mining model [25] when predicting the academic success of learners. According to this model, forecasting is carried out in sequential steps: business understanding, data understanding, data preparation, modeling, evaluation, and implementation. The effective implementation of these steps ensures the quality and integrity of the mining process and minimizes the likelihood of errors.

The rest of this document is organized as follows. Section 2 reviews related work. Section 3 describes the methodology. Section 4 presents the results. Section 5 provides conclusions and directions for further work.

2. Literature Review

Several studies have explored the use of machine learning algorithms to predict student performance in educational settings. Common classifiers considered include decision trees, random forest, naive Bayes, support vector machines, and k-nearest neighbors [26–28]. These algorithms show varying levels of accuracy, with random forests and decision trees of-

ten coming out the best. Researchers have applied these techniques to a variety of datasets, including undergraduate student records and online course data, considering factors such as grade point averages, practice exams, and written exams [26,27]. Qiu et al. [29] use classification methods for prediction and propose the e-learning performance prediction framework based on behavior classification. This system includes learning behavioral feature selection and incorporating behavioral data through feature fusion using a behavioral classification model. This process generates feature values for each behavior type category, which are then used in a machine learning-based predictor of student performance. The authors state that this method is better than traditional classification methods.

Some authors use hybrid methods to improve the prediction of student performance. Shreem et al. [30] present an innovative hybrid selection mechanism for prediction. The proposed model is a hybrid between a binary genetic algorithm, an electromagnetic-like mechanism, and k-means algorithms. The results presented demonstrate the ability of the proposed method to improve the performance of the binary genetic algorithm and the performance of all classifiers. Beckham et al. [31] use Pearson's correlation to determine which factors influence student performance and experimented with several machine learning techniques. The authors found that students are more likely to fail when they have previous failures, and another factor is the age of the student as older students fail more often than younger students. Göktepe Yıldız and Göktepe Körpeoğlu [32] explore the use of an adaptive neuro-fuzzy inference system, to model students' perceptions of their problem-solving skills based on their creative problem-solving characteristics. The findings indicated that this approach can accurately predict students' perceptions of their problem-solving skills and reveal a significant relationship between problem-solving talents and creative problem-solving features.

Another possibility explored in the literature is the use of artificial intelligence (AI) for forecasting. AI techniques such as machine learning and deep learning enable the analysis of complex patterns in behavioral data and the creation of more accurate predictive models. Baashar et al. [33] analyzed the use of neural networks to predict student performance. The findings showed that the use of artificial neural networks in combination with data analysis and data mining techniques is common practice and allows researchers to evaluate the effectiveness of their findings in assessing academic achievement. The authors noted that artificial neural networks demonstrated high accuracy in predicting the outcomes of academic achievement. However, they acknowledge that comparable results were achieved using other data mining methods. Furthermore, it was observed that the use of different data mining methods did not significantly increase the accuracy of the predictions. Cruz-Jesus et al. [34] use methods such as artificial neural networks, decision trees, extremely randomized trees, random forests, support vector machines, and k-nearest neighbors to predict academic achievement. In estimating each model, data from the beginning of each academic year were used as independent variables, and the dependent variable corresponded to the end of the year. The authors conclude that artificial intelligence methods reveal a better performance compared to traditional approaches. Recent studies have investigated various factors that influence the prediction of student performance. Both academic and non-academic parameters have been found to contribute to predictive accuracy [35]. A systematic review of machine learning models found that demographic, academic, and behavioral characteristics are commonly used for prediction, although more research is needed to generalize the results [36]. Specific factor analysis showed that exercise-related variables were the best predictors, while forum variables were less useful. Clickstream data can be effective when exercise data are not available. Prediction accuracy varies depending on the type of assignment, data collection methods, and the nature of the prediction result [37]. Yağcı [38] uses three specific parameters for prediction: mid-term exam grades, department details, and faculty details. The article highlights the importance of data-driven studies in the development of a learning analytics framework within higher education, highlighting their contribution to decision-making processes. Some authors [20,39] emphasize that self-efficacy is one of the most important elements

that allows for the prediction of academic achievements. When self-efficacy is included in psychological models that examine student academic achievement, the significance of the effect of other variables on academic achievement is reduced.

The CRISP-DM methodology is useful in educational data mining projects due to its structured and comprehensive approach. It is a standardized six-step data analysis process that includes business understanding, data understanding, data preparation, modeling, evaluation, and implementation. The effectiveness of the methodology was demonstrated in a study predicting student performance at a Croatian university, where decision tree modeling achieved a high accuracy and interpretability [40]. In addition, this methodology was used to evaluate machine learning models to predict high school student performance in the Saber 11 test in Colombia [41].

According to the literature review, this study uses the CRISP-DM data mining model as a structured prediction analysis framework, combined with classification algorithms, to increase the accuracy of academic success predictions.

3. Materials and Methods

In our study, we discuss the challenges and quality issues within higher education in relation to educational processes, and the risks of dropping out by organizing an engineering study program in a virtual learning environment. It is appropriate to analyze students' data using data mining, as data mining allows for the optimal use of big education data and extraction of useful information from them. An early-warning framework based on data mining was designed to predict the risks and academic success of learners in order to reduce the dropout percentage (see Figure 3). Learners interact with VLEs, such as Moodle, generating learning data based on their activities, outputs, and outcomes. These data, categorized into specific metrics, include overall activity (tracking clicks, login frequency, and engagement), views (tracking lectures and material views), individual tasks (tasks completed, time spent and grades received), group tasks (time and participation in collaborative assignments), tests (number of subjects passed, pass/fail rates and scores), forum participation (comments and time spent), and assessments (overall course or subject grades). These detailed metrics are collected and stored on the Moodle server, forming a training dataset for further analysis. The dataset is then used by an educational data scientist or an early-warning system server to build predictive models that estimate learners' academic success or dropout risk. These predictions are shared with ESL coordinators to target interventions for at-risk learners. The system creates a feedback loop in which interventions aim to improve learning outcomes, providing timely measures to reduce academic failure and increase student retention.

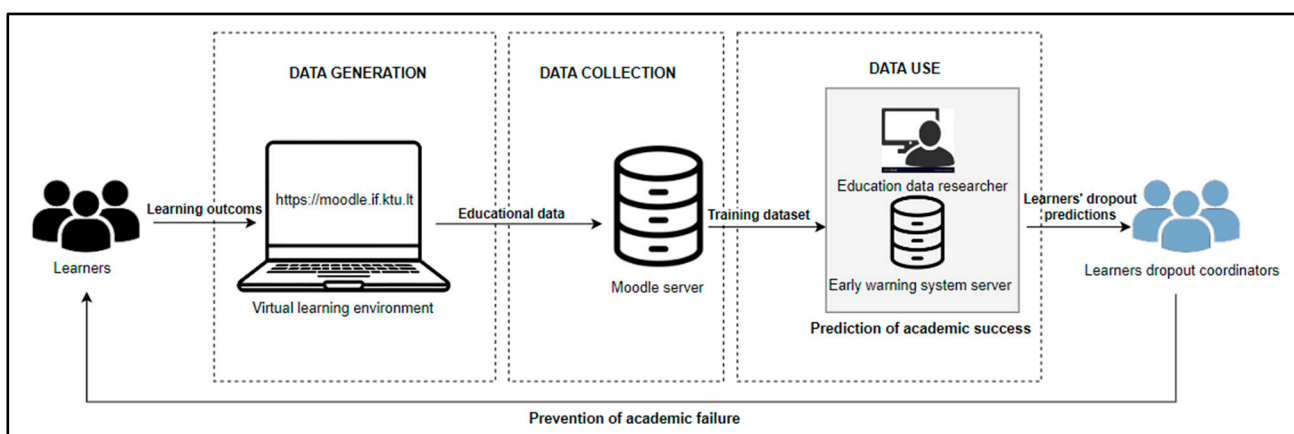


Figure 3. An early warning framework based on data mining.

The prediction of academic success was based on the CRISP-DM data mining model. The data mining software Weka 3.8 [42] was used.

3.1. Phases of the CRISP-DM Model

The CRISP-DM methodology is useful in educational data mining projects due to its structured and comprehensive approach. It is a standardized six-step data analysis process that includes business understanding, data understanding, data preparation, modeling, evaluation, and implementation [40]. The prediction was carried out according to these phases of the CRISP-DM model.

Business understanding phase. To analyze the possibilities of applying data mining to predict the academic success of “Distance Learning Information Technology” students, a SWOT analysis was performed (see Table 1).

Table 1. SWOT analysis of data mining application possibilities.

STRENGTHS	WEAKNESS
A virtual learning environment	There is a lack of information about the current situation of learners
Academic information system	Uneven assessment of learners
Experience in implementing an early warning system	Inconsistent monitoring of learner progress
Highly qualified and competent teachers	Student dropout
	Students are stressed at the end of the semester
OPPORTUNITIES	THREATS
Use personalized administrative and learning process data	Ensuring learner data protection, privacy and confidentiality
Digitize the monitoring of learners’ progress	Risk of wasting information extracted during data mining
Optimize the use of big educational data	High load on the Moodle server when retrieving data from the database
Improve the system of providing academic support	

According to SDG 4 the university pays a significant amount of attention to the quality of students’ studies. The academic achievements of students are an essential indicator of the quality of their studies, and the successful completion of the studies positively affects the reputation of the educational institution. In 2020, nine first-year students dropped out of the “Distance Learning Information Technologies” study program in the fall semester. At the university, bachelor’s studies are conducted as face-to-face studies (on-campus); therefore, the progress or attendance of students can be determined throughout the course of the semester. Master’s programs are delivered online (distance learning), so a lack of progress is noticeable only at the end of the semester. Teachers of the study program cannot identify the reasons for dropping out, because some master’s students do not even join remote lectures, do not report laboratory work, etc. For these reasons, master’s students were chosen for prediction.

Two cases are presented: The first case (1) is presented as the “Basics of Virtual Learning” and the second case (2) as “Research Project 1”. The main grades are provided only at the end of the semester, and the cumulative score can also consist of a task with a high percentage value. In this case, it is difficult to predict the learning outcomes of the student and to provide timely academic support, as the student’s academic success/failure is only known at the end of the semester when students submit/fail to submit module assignments. To predict possible student dropouts, we decided to apply the predictive model to these modules.

Data understanding and preparation phases. These steps included identifying relevant data and potential data quality issues, collecting primary data, and preparing them for the final dataset. The Moodle database stores various data about the learner’s learning progress: the learner’s login time, frequency, activities performed, grades received, etc. In the Moodle system, study program curators and teachers can receive various reports, which can be analyzed to evaluate the learning results achieved by learners, track learners’ progress, activity, etc. In addition, the university’s academic information system collects administrative data about the learner. In the virtual learning environment and the academic

information system, big educational data are collected but not analyzed by teachers. It is appropriate to analyze these data using data mining because data mining allows for the optimal use of educational data and the extraction of useful information from them.

Modeling phase. In accordance with the studies analyzed in the literature review, five classification algorithms were selected for the initial modeling stage, each with distinct advantages and limitations. (1) Decision trees are effective for many prediction problems due to their interpretability but may face challenges with smooth class boundaries. (2) The Bayesian classifier is fast, scalable, and works well with both continuous and discrete attributes, making it particularly suitable for real-time prediction scenarios. (3) The k-nearest neighbor is highly versatile, performing well in multi-class settings and with multi-labeled objects, though it can become computationally intensive when dealing with large training datasets. (4) Support vector machines are especially well-suited for binary classification problems, providing a high accuracy in distinguishing between two classes [43]. A random forest combines multiple decision trees to create a robust classifier that offers advantages such as its non-parametric nature, its ability to handle multiple data types, and its resistance to overfitting [44].

As a result of the analysis, the most suitable method for the prediction model was identified.

Evaluation and implementation phases. The model quality assessment involved the implementation of various machine learning algorithms, including the decision tree, Bayesian classifier, random forest, support vector classifier, and k-nearest neighbor classifier. An initial model was created using these algorithms, and their quality was evaluated by comparing the main evaluation metrics that highlighted the trade-off between true positive and false positive rates at different threshold values

3.2. Data Preparation

The data from the cases presented were taken from the Moodle system of the first-semester master's study modules "Basics of Virtual Learning" and "Research Project 1". Since in these modules' assessments are organized at the end of the semester (student on-time reporting and grading cannot be used as features), only two attributes were selected: (1) student logins; (2) student clicks. Structured Query Language (SQL) queries were used to extract data, which collect data on student logins and student clicks on these modules. SQL queries were first tested on a personal Moodle database running on a MySQL server. When checking the correctness of requests, the data obtained was compared immediately after its execution. Data preparation was performed on the initial data, where the information was filtered, renamed, and merged. This step resulted in a dataset to train the academic success prediction model (see Table 2).

Table 2. Structure of the training dataset for the academic success prediction model.

Variable	Value
TP1_access_week	The number of student logins to the "Research Project 1" module
TP1_clicks_week	The number of students clicks within the "Research Project 1" module
VMP_access_week	The number of student logins to the "Basics of Virtual Learning" module
VMP_clicks_week	The number of students clicks in the "Basics of Virtual Learning" module
key	Student identity pseudonymization key (125 students)
success	A class variable with F representing academic failure and T representing academic success

To ensure the protection of the student's data, privacy, and confidentiality, first, the data are pseudonymized and a key is created for each student's data (see Figure 4). The key protects the identification of the learners while developing a model to predict academic success, and once the model is developed and implemented, the key allows the study program administration to identify struggling learners.

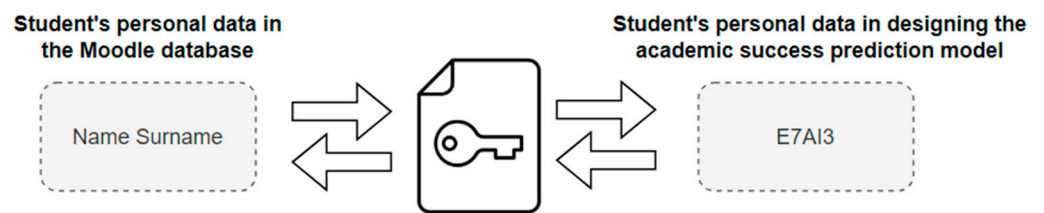


Figure 4. Pseudonymization of student data.

The prediction was carried out in several stages, considering changes in results, with data for 5 weeks, data for 6 weeks, data for 7 weeks, and data for 8 weeks. According to Ortiz-Lozano et al. [45], the initial year of studies, particularly the first 6–7 weeks, is considered to be significant for the prevention of academic failure.

3.3. Modeling the Prediction of Academic Success

The decision tree algorithm, Bayesian classifier, random forest algorithm, support vector classifier, and k-nearest neighbors' classifier were selected for modeling. An initial model was used to evaluate the quality results of the algorithms, and the following parameters were compared: Precision, Recall, F-Measure, and ROC (Receiver Operating Characteristic) (see Table 3) [46].

Table 3. Performance of algorithms in an initial model for predicting academic success.

Algorithm	Data	Precision	Recall	F-Measure	ROC Area	Class
Decision tree	5 weeks	0.5	0.278	0.357	0.531	F
		0.752	0.888	0.814	0.531	T
	6 weeks	0.556	0.417	0.476	0.676	F
		0.786	0.865	0.824	0.676	T
	7 weeks	0.441	0.417	0.429	0.592	F
		0.769	0.787	0.778	0.592	T
	8 weeks	0.429	0.417	0.423	0.598	F
		0.767	0.775	0.771	0.598	T
Bayesian classifier	5 weeks	0.361	0.611	0.454	0.646	F
		0.781	0.562	0.654	0.646	T
	6 weeks	0.418	0.778	0.544	0.716	F
		0.862	0.562	0.68	0.716	T
	7 weeks	0.41	0.694	0.515	0.737	F
		0.828	0.596	0.693	0.737	T
	8 weeks	0.414	0.667	0.511	0.735	F
		0.821	0.618	0.705	0.735	T
Random forest	5 weeks	0.9	0.25	0.391	0.648	F
		0.765	0.989	0.863	0.648	T
	6 weeks	0.857	0.333	0.48	0.734	F
		0.784	0.978	0.87	0.734	T
	7 weeks	0.824	0.389	0.528	0.772	F
		0.796	0.966	0.873	0.772	T
	8 weeks	0.765	0.361	0.491	0.716	F
		0.787	0.955	0.863	0.716	T
Support vector classifier	5 weeks	0.556	0.139	0.222	0.547	F
		0.733	0.955	0.829	0.547	T
	6 weeks	0.692	0.25	0.367	0.603	F
		0.759	0.955	0.846	0.603	T
	7 weeks	0.737	0.389	0.509	0.666	F
		0.792	0.944	0.862	0.666	T
	8 weeks	0.737	0.389	0.509	0.666	F
		0.792	0.944	0.862	0.666	T

Table 3. Cont.

Algorithm	Data	Precision	Recall	F-Measure	ROC Area	Class
K-nearest neighbors classifier	5 weeks	0.381	0.444	0.41	0.561	F
		0.759	0.708	0.733	0.561	T
	6 weeks	0.357	0.417	0.385	0.539	F
		0.747	0.697	0.721	0.539	T
	7 weeks	0.425	0.472	0.447	0.613	F
		0.776	0.742	0.759	0.613	T
	8 weeks	0.5	0.5	0.5	0.655	F
		0.798	0.798	0.798	0.655	T

Table 3 shows the classification *Precision*, *Recall*, *F-Measure*, and *ROC* results. *Precision* was calculated as the number of true positives divided by the total number of positive and negative observations (see Formula (1)).

$$Precision = True\ Positive / (True\ Positive + False\ Positive), \quad (1)$$

The result is represented as a value ranging from 0.0, indicating no accuracy, to 1.0, indicating complete or perfect accuracy. *Recall* calculates the proportion of correctly predicted positive instances to all possible positive predictions within the dataset. This metric can range from 0.0, indicating no recall at all, to 1.0, indicating complete or perfect recall. *F-Measure* provides the ability to combine *Precision* and *Recall* into a single metric that captures both properties. The *F-Measure* is calculated as follows:

$$F-Measure = (2 * Precision * Recall) / (Precision + Recall), \quad (2)$$

A low *F-Measure* score is 0.0, indicating a poor performance, while a high or perfect *F-Measure* score is 1.0. The *ROC* value is useful for determining the ability of a model to discriminate between classes [47].

When examining the results of the correct predictions of the algorithms, it is evident that the random forest algorithm provided high values for the parameters considered in all the data instances considered, compared to other algorithms. The random forest algorithm achieved the highest results using seven weeks' worth of data, correctly predicting 80% of the cases. Comparing the precision of the algorithms over the entire period in both classes, the precision of the random forest algorithm was 81%, the support vector classifier was 72%, the decision tree was 63%, the Bayesian classifier was 61%, and the k-nearest neighbors' classifier was 59%. The random forest algorithm has also achieved the highest value of *F-measure* (0.873) among all the algorithms evaluated. In the results of this algorithm, the value of the *F-Measure* was the highest with data for the entire period compared to the other algorithms. The support vector classifier also shows high *F-Measure* results of 0.862 (7–8 weeks), 0.846 (6 weeks), and 0.829 (5 weeks), respectively.

Based on the results obtained, it can be concluded that the quality parameters of the random forest and support vector classifier are better than those of other applied algorithms. Comparing the results of these algorithms with the 7-week data, it can be concluded that the random forest algorithm is superior to the support vector classifier in predicting academic success (by assigning a value to T).

A final model for predicting academic success was created using a random forest algorithm (see Figure 5).

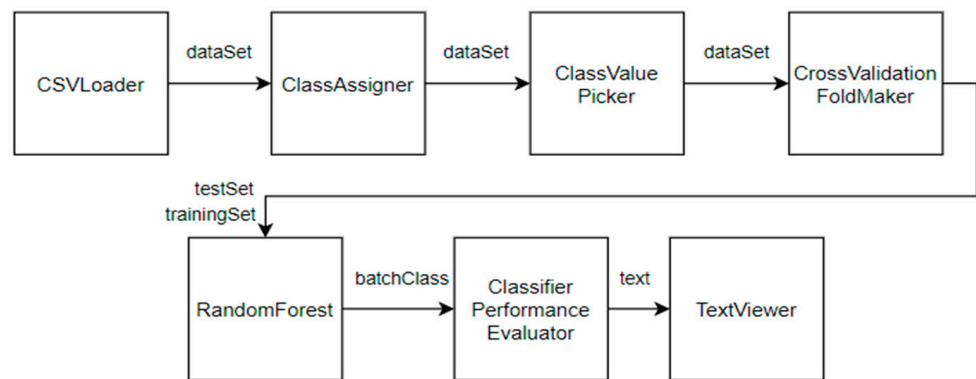


Figure 5. A model for predicting student academic success.

The academic success prediction model consists of the following seven components: (1) a “CSVLoader” component designed to load a dataset in CSV (.csv) format; (2) a “ClassAssigner” component that specifies the index of a class variable (in this case, the variable “success” whose index is “last”); (3) a “ClassValuePicker” component that specifies the value of a class variable (in this case the value “N”, which is “/first”); (4) a “CrossValidationFoldMaker” component that specifies how many times and into how many parts the dataset is split (in this case, part for training data and part for testing); (5) a “RandomForest” component that indicates that a random forest algorithm is applied to the model; (6) a “ClassifierPerformanceEvaluator” component for generating prediction results; and (7) a “TextViewer” component for viewing the results in a text format.

4. University Case Study on Predicting Academic Performance

To assess the suitability of the developed model, two tests were conducted: one involved testing SQL queries on the Moodle database, while the other focused on testing the accuracy of the academic success prediction model.

SQL SELECT queries were prepared and used to retrieve data from the Moodle database. These queries were written on a personal database server (server specifications: macOS X, Apache (2.2.23), PHP (7.4.2), MySQL (5.7.26)), and testing was performed on the Moodle database (server specifications: (Debian Linux 10, Apache (2.4.38), PHP (7.4.33), MariaDB (10.4.28)). The accuracy of the SQL queries was considered during the testing process, ensuring that they were free from syntax errors and providing the correct data from the Moodle database.

The test was carried out using data from students enrolled in 2021 and 2022. The number of logins to the modules “Basics of Virtual Learning” and “Research Project 1” and clicks within these modules were checked. The academic success prediction model was tested using three different datasets: (1) a dataset prepared for model training; (2) a dataset with data from students enrolled in 2021 (21 students); and (3) a dataset with data from enrolled students enrolled in 2022 (74 students). The results of the prediction of the academic success of students in 2021 are presented in Figure 6.

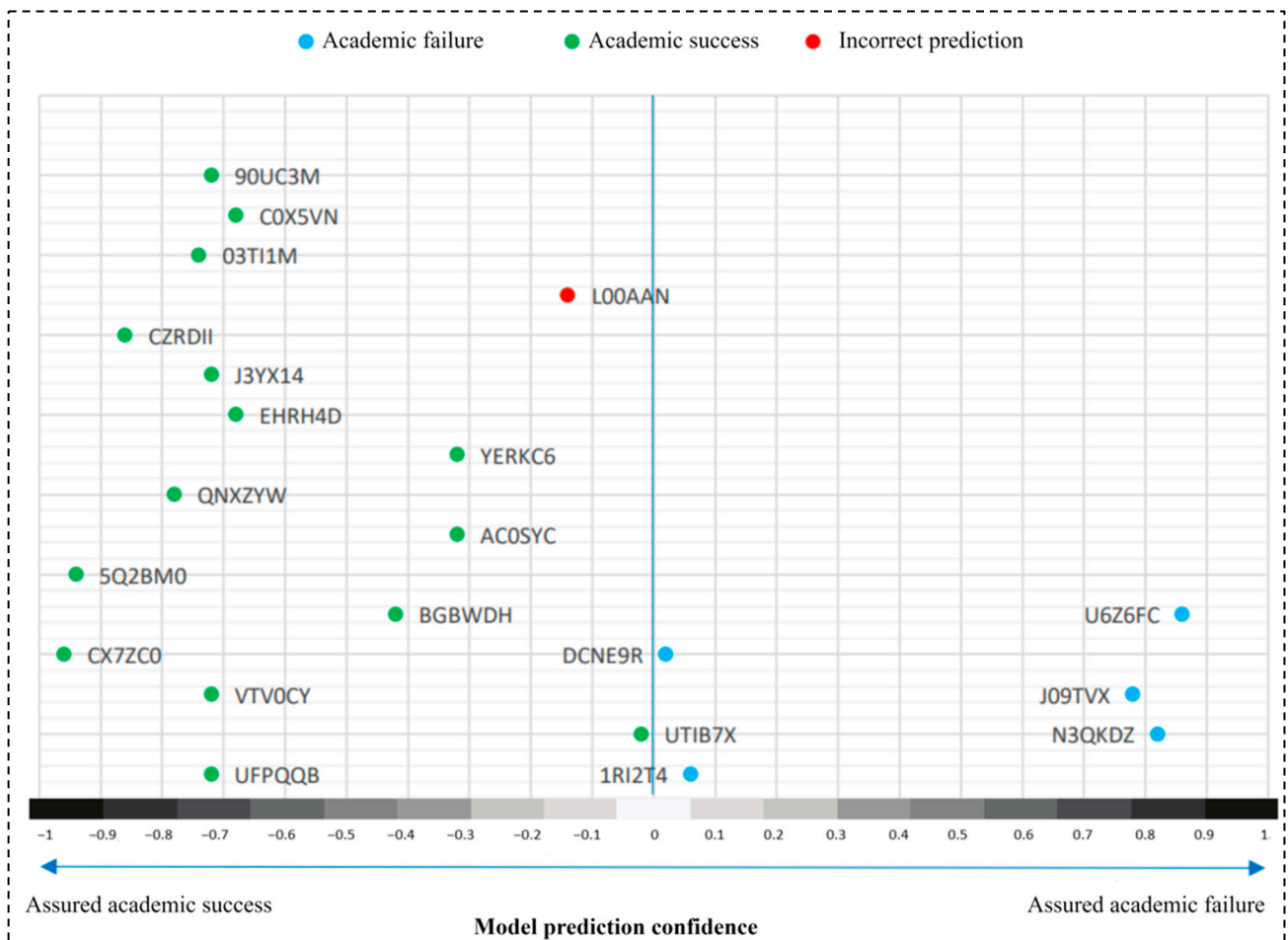


Figure 6. Academic success prediction for students enrolled in 2021.

The model predicts that 25% of students are at risk of not completing their studies. For these students, the model assigned a value of F. Based on the confidence values, it can be stated that the model's prediction of the academic failure of the two students is uncertain as the confidence level obtained is less than 0.1. The prediction of the academic success of two other students is also unlikely, with a confidence interval of less than -0.2 .

The results of the prediction of student academic success in 2022 are presented in Figure 7.

Figure 7 shows that 14% (10 out of 74) of the students are at risk of dropping out. They were assigned a value of F. In this case, the confidence of the model was weak when setting one student at the value F, and the model evaluated that the possibility of one other student stopping or continuing their studies was equal (confidence estimate equal to 0). Considering the confidence of the model, assigning a value to T identifies five students whose confidence estimates were less than -0.2 .

The results were compared with real data on the learning situations of students in 2021 and 2022. Based on this, the errors made by the model are visualized as incorrect predictions in Figures 6 and 7. Comparing the predictions provided by the model and information about the real situation, it can be concluded that the model correctly assigned the value of F in 73% of cases, i.e., 11 out of the 15 students predicted to drop out did so. It is important to note that in cases where the model was uncertain (five cases) or incorrect (three cases), the students showed signs of academic failure, including academic debt, low academic achievement, or absenteeism. The results also showed that in 81% of the cases the model correctly predicted that the students would stay in their studies. Unfortunately,

15 of the 80 students who were predicted to continue their studies (assigned a T value) dropped out or went on academic leave for various personal reasons.

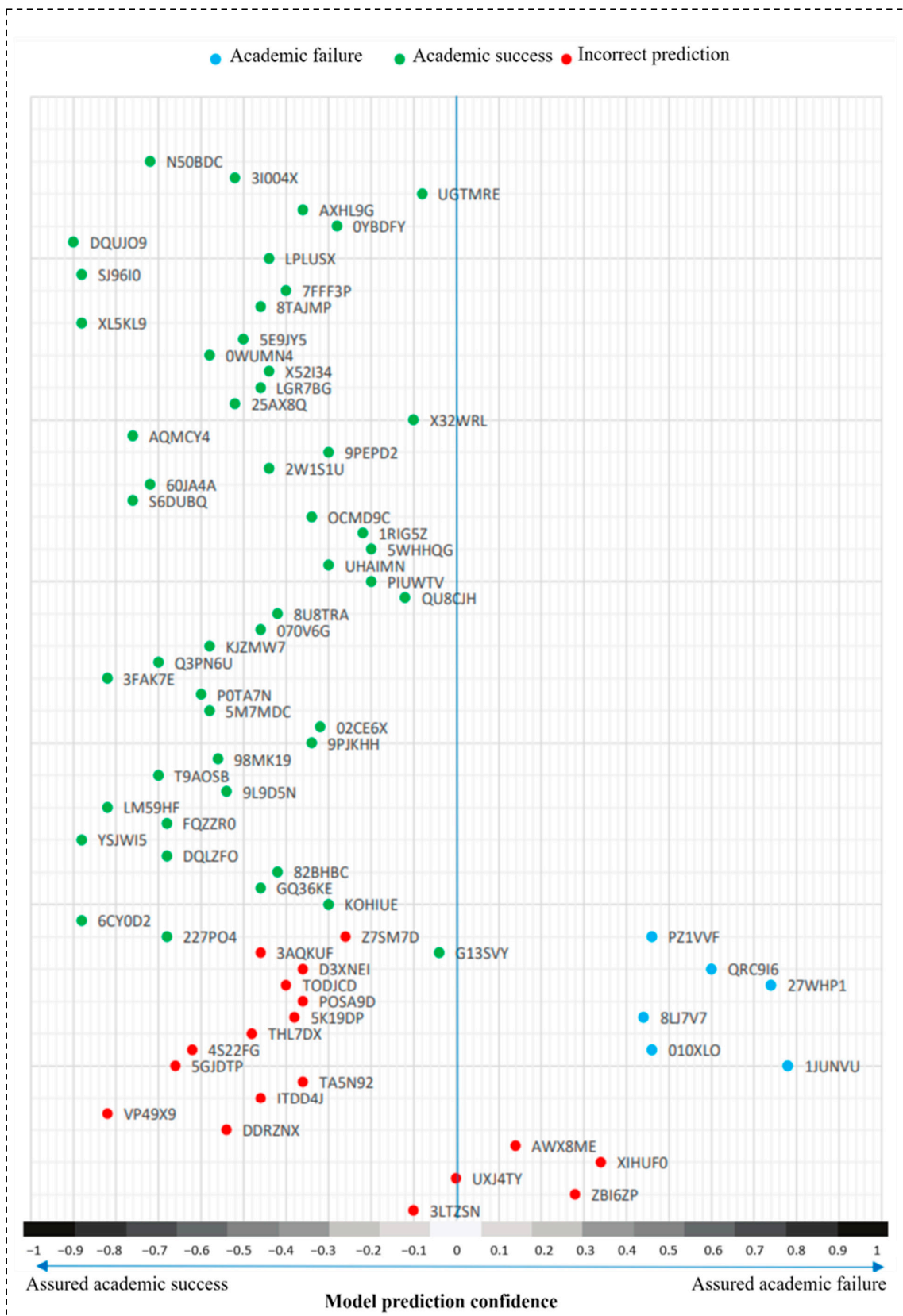


Figure 7. Academic success predictions for students enrolled in 2022.

However, in general, the results obtained revealed that by using the data of the student learning process collected by the virtual learning environment and applying data mining to their analysis, it is possible to predict which students are at risk of dropping out. Random forest is an effective machine learning technique to predict student dropout, offering advantages such as robustness to outliers and noise, estimation of importance of characteristics, and high accuracy [48]. Other studies have demonstrated its effectiveness, with accuracy rates ranging from 73% to 87.6%, along with other strong performance measures such as precision, recall, and F1 score [48,49].

5. Conclusions

In this study, we suggest applying data mining and classification algorithms to predict academic success. A random forest algorithm was chosen for the model based on the results of the primary analysis of the various algorithms. The CRISP-DM data mining model was used to predict the academic success of the learners, allowing the prediction to be carried out in successive stages. This study contributes to the integration of data mining methods with a focus on predicting academic risk in specific study programs. The originality of this study lies in its focus not only on identifying at-risk students, but also on opportunities to communicate this risk to students to improve retention rates.

The findings highlight that, after analyzing the learning process data with the proposed academic success prediction model, it is possible to identify students who are at risk of dropping out. These results are consistent with previous research that highlights the utility of machine learning algorithms such as random forests in educational data mining. For example, previous studies [44,48] have demonstrated the effectiveness of similar methods in identifying at-risk learners. However, this study advances this field by incorporating the CRISP-DM methodology, which increases model reliability and interpretability. However, it would be reasonable to improve the model to reduce the probability of errors and increase the accuracy of the prediction.

The main limitation of the proposed model is as follows: the module returns some incorrect values during the prediction. Despite this limitation, the proposed model can help identify potential academic failures over time and ensure sustainable education.

Future work will include supplementing the early warning model with an assessment of students' academic self-efficacy, which that would be administered during the introductory week of study. The dataset created for model training should be supplemented annually with new data on students who have completed/discontinued their studies. Furthermore, we plan to supplement the prediction model with other study modules and features, such as including a module with earlier grading opportunities (not just at the end of the semester) and include evaluations of work from throughout the semester as a feature. Such improvements could potentially reduce the number of incorrect prediction values.

Author Contributions: Conceptualization, E.S., D.G. and A.K.; methodology, E.S., D.G., V.P. and A.K.; formal analysis, E.S., D.G. and A.K.; writing—original draft preparation, E.S., D.G., V.P. and A.K.; writing—review and editing, E.S. and D.G.; visualization, E.S. and A.K.; project administration, E.S.; funding acquisition, D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study, as this study involves no more than minimal risk to subjects.

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

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Srivastava, J.; Srivastava, A.K. Understanding linkage between data mining and statistics. *Int. J. Eng. Technol. Manag. Appl. Sci.* **2015**, *3*, 4–12.
2. Manjarres, A.V.; Sandoval, L.G.M.; Suárez, M.S. Data mining techniques applied in educational environments: Literature review. *Digit. Educ. Rev.* **2018**, *33*, 235–266. [[CrossRef](#)]
3. Batool, S.; Rashid, J.; Nisar, M.W.; Kim, J.; Kwon, H.Y.; Hussain, A. Educational data mining to predict students' academic performance: A survey study. *Educ. Inf. Technol.* **2023**, *28*, 905–971. [[CrossRef](#)]
4. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1355. [[CrossRef](#)]
5. Hernández-Blanco, A.; Herrera-Flores, B.; Tomás, D.; Navarro-Colorado, B. A systematic review of deep learning approaches to educational data mining. *Complexity* **2019**, *2019*, 1306039. [[CrossRef](#)]
6. Behr, A.; Giese, M.; Tegum Kamdjou, H.D.; Theune, K. Dropping out of university: A literature review. *Rev. Educ.* **2020**, *8*, 614–652. [[CrossRef](#)]
7. Alyahyan, E.; Düşteğör, D. Predicting academic success in higher education: Literature review and best practices. *Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 3. [[CrossRef](#)]
8. Trakunphutthirak, R.; Cheung, Y.; Lee, V.C. A study of educational data mining: Evidence from a thai university. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 734–741. [[CrossRef](#)]
9. Rajalaxmi, R.R.; Natesan, P.; Krishnamoorthy, N.; Ponni, S. Regression model for predicting engineering students academic performance. *Int. J. Recent Technol. Eng.* **2019**, *7*, 71–75.
10. Križanić, S. Educational data mining using cluster analysis and decision tree technique: A case study. *Int. J. Eng. Bus. Manag.* **2020**, *12*, 1847979020908675. [[CrossRef](#)]
11. Chen, J.; Zhao, J. An Educational Data Mining Model for Supervision of Network Learning Process. *Int. J. Emerg. Technol. Learn.* **2018**, *13*, 67. [[CrossRef](#)]
12. Doko, E.; Bexheti, L.A.; Hamiti, M.; Etemi, B.P. Sequential Pattern Mining Model to Identify the Most Important or Difficult Learning Topics via Mobile Technologies. *Int. J. Interact. Mob. Technol.* **2018**, *12*, 109–122. [[CrossRef](#)]
13. Paiva, R.; Bittencourt, I.I.; Lemos, W.; Vinicius, A.; Dermeval, D. Visualizing learning analytics and educational data mining outputs. In *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part II 19* (pp. 251–256); Springer International Publishing: New York, NY, USA, 2018. [[CrossRef](#)]
14. Almasri, A.; Alkhalaf, R.S.; Çelebi, E. Clustering-based EMT model for predicting student performance. *Arab. J. Sci. Eng.* **2020**, *45*, 10067–10078. [[CrossRef](#)]
15. Khasanah, A.U. A comparative study to predict student's performance using educational data mining techniques. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2017; Volume 215, p. 012036. [[CrossRef](#)]
16. Seidel, E.; Kutieleh, S. Using predictive analytics to target and improve first year student attrition. *Aust. J. Educ.* **2017**, *61*, 200–218. [[CrossRef](#)]
17. Arulkadacham, L.; McKenzie, S.; Aziz, Z.; Chung, J.; Dyer, K.; Holt, C.; Mundy, M. General and unique predictors of student success in online courses: A systematic review and focus group. *J. Univ. Teach. Learn. Pract.* **2021**, *18*, 7. [[CrossRef](#)]
18. Yokoyama, S. Academic self-efficacy and academic performance in online learning: A mini review. *Front. Psychol.* **2019**, *9*, 2794. [[CrossRef](#)]
19. Doménech-Betoret, F.; Abellán-Roselló, L.; Gómez-Artiga, A. Self-efficacy, satisfaction, and academic achievement: The mediator role of Students' expectancy-value beliefs. *Front. Psychol.* **2017**, *8*, 1193. [[CrossRef](#)]
20. Nasir, M.; Iqbal, S. Academic Self Efficacy as a Predictor of Academic Achievement of Students in Pre Service Teacher Training Programs. *Bull. Educ. Res.* **2019**, *41*, 33–42.
21. Quinn, R.J.; Gray, G. Prediction of student academic performance using Moodle data from a Further Education setting. *Ir. J. Technol. Enhanc. Learn.* **2020**, *5*, 1–19. [[CrossRef](#)]
22. Hellas, A.; Ihantola, P.; Petersen, A.; Ajanovski, V.V.; Gutica, M.; Hynninen, T.; Liao, S.N. Predicting academic performance: A systematic literature review. In Proceedings of the Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, Larnaca, Cyprus, 2–4 July 2018; pp. 175–199. [[CrossRef](#)]
23. Yildiz, M.; Börekci, C. Predicting Academic Achievement with Machine Learning Algorithms. *J. Educ. Technol. Online Learn.* **2020**, *3*, 372–392. [[CrossRef](#)]
24. Phauk, S.; Okazaki, T. Integration of Educational Data Mining Models to a Web-Based Support System for Predicting High School Student Performance. *Int. J. Comput. Inf. Eng.* **2021**, *15*, 131–144.
25. Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. CRISP-DM 1.0: Step-by-step data mining guide. *Cris. Consort* **2000**, *76*. Available online: <https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf> (accessed on 25 November 2024).
26. Khairy, D.; Alharbi, N.; Amasha, M.A.; Areed, M.F.; Alkhalaf, S.; Abougalala, R.A. Prediction of student exam performance using data mining classification algorithms. *Educ. Inf. Technol.* **2024**, *29*, 21621–21645. [[CrossRef](#)]

27. Al Nagi, E.; Al-Madi, N. Predicting students performance in online courses using classification techniques. In Proceedings of the 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), Valencia, Spain, 19–22 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 51–58.
28. NANuradha, C.; Velmurugan, T. A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. *Indian J. Sci. Technol.* **2015**, *8*, 1–12.
29. Qiu, F.; Zhang, G.; Sheng, X.; Jiang, L.; Zhu, L.; Xiang, Q.; Chen, P.K. Predicting students' performance in e-learning using learning process and behaviour data. *Sci. Rep.* **2022**, *12*, 453. [[CrossRef](#)]
30. Shreem, S.S.; Turabieh, H.; Al Azwari, S.; Baothman, F. Enhanced binary genetic algorithm as a feature selection to predict student performance. *Soft Comput.* **2022**, *26*, 1811–1823. [[CrossRef](#)]
31. Beckham, N.R.; Akeh, L.J.; Mitaart, G.N.P.; Moniaga, J.V. Determining factors that affect student performance using various machine learning methods. *Procedia Comput. Sci.* **2023**, *216*, 597–603. [[CrossRef](#)]
32. Göktepe Yıldız, S.; Göktepe Körpeoğlu, S. Prediction of students' perceptions of problem solving skills with a neuro-fuzzy model and hierarchical regression method: A quantitative study. *Educ. Inf. Technol.* **2023**, *28*, 8879–8917. [[CrossRef](#)]
33. Baashar, Y.; Alkaws, G.; Mustafa, A.; Alkahtani, A.A.; Alsariera, Y.A.; Ali, A.Q.; Tiong, S.K. Toward predicting student's academic performance using artificial neural networks (ANNs). *Appl. Sci.* **2022**, *12*, 1289. [[CrossRef](#)]
34. Cruz-Jesus, F.; Castelli, M.; Oliveira, T.; Mendes, R.; Nunes, C.; Sa-Velho, M.; Rosa-Louro, A. Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. *Heliyon* **2020**, *6*, e04081. [[CrossRef](#)]
35. Aggarwal, D.; Mittal, S.; Bali, V. Significance of non-academic parameters for predicting student performance using ensemble learning techniques. *Int. J. Syst. Dyn. Appl.* **2021**, *10*, 38–49. [[CrossRef](#)]
36. Balaji, P.; Alelyani, S.; Qahmash, A.; Mohana, M. Contributions of machine learning models towards student academic performance prediction: A systematic review. *Appl. Sci.* **2021**, *11*, 10007. [[CrossRef](#)]
37. Moreno-Marcos, P.M.; Pong, T.C.; Munoz-Merino, P.J.; Kloos, C.D. Analysis of the factors influencing learners' performance prediction with learning analytics. *IEEE Access* **2020**, *8*, 5264–5282. [[CrossRef](#)]
38. Yağcı, M. Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ.* **2022**, *9*, 11. [[CrossRef](#)]
39. Honicke, T.; Broadbent, J. The influence of academic self-efficacy on academic performance: A systematic review. *Educ. Res. Rev.* **2016**, *17*, 63–84. [[CrossRef](#)]
40. Oreški, D.; Zamuda, D. Machine Learning Based Model for Predicting Student Outcomes. In Proceedings of the 12th International Conference on Industrial Engineering and Operations Management (IEOM 2022), Istanbul, Turkey, 7–10 March 2022; pp. 4884–4894.
41. Solano, J.A.; Cuesta, D.J.L.; Ibáñez, S.F.U.; Coronado-Hernández, J.R. Predictive models assessment based on CRISP-DM methodology for students performance in Colombia-Saber 11 Test. *Procedia Comput. Sci.* **2022**, *198*, 512–517. [[CrossRef](#)]
42. Weka Wiki Homepage. Available online: https://waikato.github.io/weka-wiki/downloading_weka/ (accessed on 27 May 2023).
43. Deeba, K.; Amutha, B. Classification algorithms of data mining. *Indian J. Sci. Technol.* **2016**, *9*. [[CrossRef](#)]
44. Horning, N. Random Forests: An algorithm for image classification and generation of continuous fields data sets. In Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan, 9–11 December 2010; Volume 911, pp. 1–6.
45. Ortiz-Lozano, J.M.; Rua-Vieites, A.; Bilbao-Calabuig, P.; Casadesús-Fa, M. University student retention: Best time and data to identify undergraduate students at risk of dropout. *Innov. Educ. Teach. Int.* **2018**, *57*, 74–85. [[CrossRef](#)]
46. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061.
47. Bouckaert, R.R.; Frank, E.; Hall, M.; Kirkby, R.; Reutemann, P.; Seewald, A.; Scuse, D. *WEKA Manual for Version 3-8-3*; University of Waikato: Hamilton, New Zealand, 2018; pp. 1–327.
48. Dass, S.; Gary, K.; Cunningham, J. Predicting student dropout in self-paced MOOC course using random forest model. *Information* **2021**, *12*, 476. [[CrossRef](#)]
49. Jayaraman, J. Predicting student dropout by mining advisor notes. In Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020), Virtual, 10–13 July 2020; pp. 629–632.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.