Predictive methods of Learning Analytics and Educational Data Mining in higher education based on Machine Learning algorithms: A Systematic Literature Review

Vanja Čotić Poturić
University of Rijeka, Faculty of Informatics and Digital Technologies
Radmile Matejčić 2, 51000 Rijeka, Croatia
e-mail: vcotic@uniri.hr

Abstract. Learning and teaching, student progress, analyzing educational data, designing assessments, and using evidence to improve learning and teaching are the subject of many research. Two areas address these issues, Learning Analytics and Educational Data Mining. Both areas have the same goal of improving the teaching and learning process and use similar techniques and methods in processing educational data from e-learning systems such as classification, grouping, regression and visualization. This article presents a systematic literature review of the last five years on predictive methods of Learning Analytics and Educational Data Mining based on Machine Learning algorithms to examine the area and provide recommendations for future research.

Keywords: Learning Analytics, Educational Data Mining, prediction, Machine Learning, systematic literature review.

1 Introduction

Learning and teaching, student progress, analyzing educational data, designing assessments, and using evidence to improve learning and teaching are the subject of many research. In the education system, Learning Management Systems (LMS) are used in various forms of teaching; in classic teaching supplemented by information and communication technologies, in hybrid learning and in distance learning. LMS combine a set of functionalities that allow teacheres to carry out activities in an online environment (delivery of learning materials, communication with students, organization of e-activities, assessment) [1].

LMS provide data about student activity as click-based (data that describes whether, when, and how often students access resources that provide different views of content), and data that reflects student actions in the course (such as participation in discussion forums and completion of assignments). This data is available in data-driven reports embedded in the LMS, but these reports are often primarily descriptive, telling participants what happened but not why it happened, and do not predict outcomes or advise participants on how to improve their results.

These problems are dealt with by two areas, Learning Analytics (LA) and Educational Data Mining (EDM). Both areas are interdisciplinary, and the main disciplines involved in these areas are shown in Figure 1.

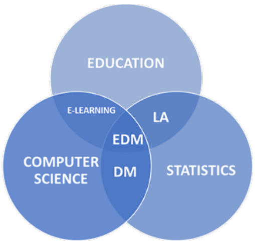

Figure 1. Main areas of EDM and LA. Source: Romero & Ventura [2].

Also, both areas have been developing rapidly over the past decade, as shown in Figure 2.

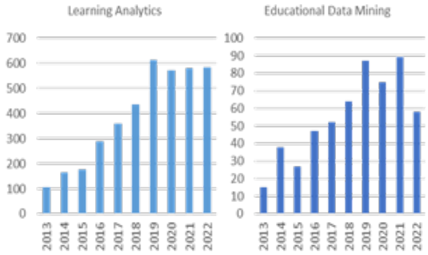

Figure 2. Number of articles in the Scopus database for the search "Learning Analytics" and "Educational Data Mining" in the last 10 years.

In the following, we will first discuss in more detail in section 2 the areas of Learning Analytics and Educational Data Mining that are the subject of this research, as well as the Predictive Modeling and Machine Learning algorithms used for predictions in these areas. Section 3 reviews the literature on research in Learning Analytics and Educational Data Mining. Then, section 4 presents the methodology chosen for the literature review. Section 5 presents the results of the systematic literature review with respect to the defined research questions, while section 6 discusses these results. Finally, section 7 concludes this paper and provides recommendations for future research.

2 Representing the areas

LA and EDM share the same goal of enhancing the teaching and learning process by improving the assessment process, understanding educational problems, and planning interventions [3], using similar techniques and methods such as classification, grouping, regression, and visualization.

Common methods of Learning Analytics and Educational Data Mining are listed in Table 1.

Table 1. Some of the usual LA and EDM methods. Source: Adapted from Romero&Ventura [2], Liñán&Pérez [4].

Method	Description	Application
Prediction	Predicting the value of the target variable from the known values of the other variables.	Prediction of student performance.
Grouping	Identify groups of similar observations.	Grouping students based on their learning patterns.
Relationship mining	To study relationships between variables and encode rules.	Identifying relationships in student behavior patterns and diagnosing difficulties.
Discovering outstanders	Point out significantly different individuals.	Detection of students with difficulties.
Social network analysis	Analyze social relationships between entities in networked information.	Interpretation of structure and relationships in collaborative activities.
Text mining	Extract high-quality information from text.	Analysis of forum content, documents, web pages.
Factorization of a non-negative matrix	Define a matrix M of positive numbers representing test result data that can be decomposed into two matrices: matrix Q representing the matrix of items and matrix S representing the student's mastery of skills.	Assessment of student skills.

The subject of this review is the first method listed in the table, prediction, the aim of which is to predict the values of the target variable from the known values of the other variables.

According to Siemens and Baker [3] it is possible to identify five main differences between EDM and LA which are listed in Table 2.

Table 2. Key differences between EDM and LA. Source: Adapted from Siemens & Baker [3]

Characteristics	LA	EDM
Discovery	goal is to exploit human judgment	automated discovery, using human judgment is a tool for this
Reduction	wants to understand whole systems	reduces systems to components and explores them and their relationships
Origin	semantic web, intelligent curriculum, outcome prediction, systematic interventions	educational software, modeling of students
Adjusment and personalization	informs and empowers students and professors	automated adaptation
Techniques and methods	social network analysis, analysis of feelings, influence, discourse, prediction of success	classification, clustering, Bayesian modeling, relationship mining, visualization

It can be said that EDM focuses more on techniques and methods, while LA is more about application. However, these differences are less noticeable as both areas evolve over time [4].

2.1 Educational Data Mining

Data Mining (DM) is the extraction of hidden useful information from a data set through scientific analysis and methods that identify data trends and hidden patterns within a given data set, and as such Data Mining can be characterized as knowledge discovery [5, 6].

Data Mining can be applied in various fields, one of them is education. Data Mining applied to educational data is referred to as Educational Data Mining (EDM) [7]. A popular definition for the field of EDM is proposed by the International Society for Data Mining in Education in 2018: "EDM is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in."

This area deals with the development of methods that discover knowledge from data of the educational environment [8], including Machine Learning, Psychometrics and other areas of Statistics, Information Visualization and Computer Modeling [9] (Figure 1).

EDM is used to identify learning challenges, examine, and predict student performance [10–12] and to assess the integration of technology into the learning process [13].

The three most common problems for which predictive DM methods are used in education are to find out whether a student will pass or fail a certain course [14], to predict the grades of a certain exam or final grades [15] and to identify those students who are at risk to drop out [16].

2.2 Learning Analytics

Learning Analytics includes the measurement, data collection, analysis and reporting (visualization) of data about students and learners in general [17]. The 2011 International Conference on Learning Analytics LAK adopted the following definition of this area: “LA is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs.”

Three key elements appear in this definition: data, analysis and action (Figure 3) [18]. Data is a set of collected information about the learner, the learning environment, learning interactions and learning outcomes. This information is usually collected during the learning process. Data analysis is a process that provides insights into actions that can be taken on data. It is based on a set of mathematical and statistical algorithms. Machine Learning algorithms can also be used in this step. Taking action is the ultimate goal of every Learning Analytics to improve the learning and teaching process.

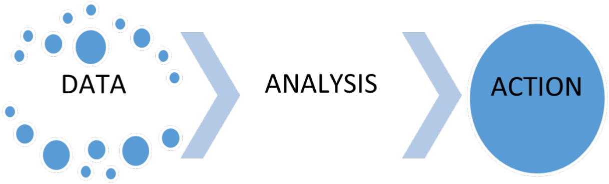

Figure 3. Three key elements of Learning Analytics.

Learning Analytics is used to understand and optimize learning and teaching, to quickly identify risks for students, to propose interventions to help students learn, to detect weaknesses in the education system and to improve it, etc. [19].

The LA area combines (Figure 1) Education (educational research, educational technologies), Analytics (Computer/Data science, Artificial Intelligence, Statistics, Visualization) and human-centered design (participatory design, usability) [20]. This field of research has developed rapidly in the last decade (Figure 2).

Initially, the most common use of Learning Analytics was to predict student academic performance, specifically to identify students at risk of failing or dropping out of courses. Today, more productive and powerful methods of using learning analytics to support teaching and learning are being implemented.

Some of the current goals of Learning Analytics are to help students develop lifelong learning skills and strategies, to provide students with personalized and timely feedback on their learning, to support the development of important skills such as collaboration, critical thinking, communication and creativity, to support quality learning and teaching by providing empirical evidence of the success of pedagogical innovations.

There are four main categories of Learning Analytics (Figure 4). Descriptive answers the question of what happened, diagnostic tells why it happened, predictive answers the question of what will happen next, while prescriptive tells what needs to be done for improvement [21].

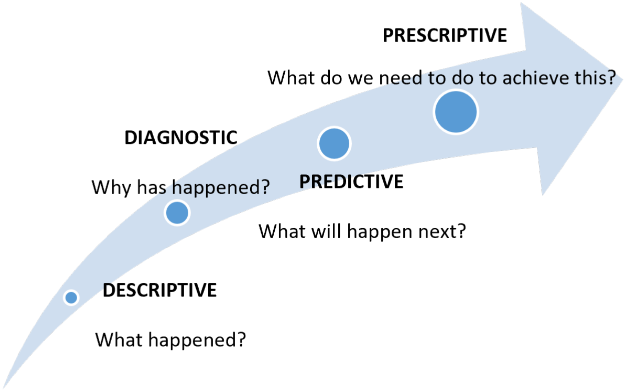

Figure 4. Learning Analytics categories.

2.3 Predictive Modeling

In both fields, Educational Data Mining and Learning Analytics, Predictive Modeling has become a core practice of researchers, with a major focus on predicting student success [22].

Predictive Modeling is a group of techniques used to draw conclusions about uncertain future events. In education, we may have value in predicting measures of learning (e.g., student academic performance or skill acquisition), teaching (e.g., impact of a particular teaching style or teacher on an individual), or other proxy metrics for organizations (e.g., predicting course retention or enrollment).

The purpose of Predictive Modeling is to create a model that predicts the values (or class if prediction is not concerned with numeric data) of new data based on observations. It is based on the assumption that a set of known data can be used to predict the value or class of new data based on the observed variables.

In Predictive Modeling, a hold-out dataset is used to assess the model’s suitability for prediction and to protect against overfitting the model to the data used for training. There are various strategies for creating hold-out dataset (model validation), including k-fold cross-validation, leave-one-out cross-validation, random subsampling, etc.

The steps of Predictive Modeling are problem identification, data collection, feature engineering, feature selection, model building, and model evaluation. The main characteristics of each step are listed in Table 3.

Table 3. Steps of Predictive Modeling.

Problem identification	It is necessary to select a problem that will repeat itself in the future, in which there are measurable characteristics, a clear outcome, the possibility of intervention and a large set of data.
Data collection	The researcher needs to identify the output variable (eg final grade or achievement level) as well as possible input variables (eg gender, previous level grade, number of LMS accesses). Data types differ according to the information content of their measurement scales. Categorical data, which include nominal and ordinal data, allow only qualitative classification into certain categories. It is possible to perform different mathematical operations on numerical data, depending on whether it is intervals or ratios.
Feature engineering	Feature engineering is a preprocessing step that transforms raw data into variables that can be used in predictive models.
Feature selection	In order to build and apply a predictive model, it is necessary to select predictive (input) variables that are correlated with the output variable, the value to be predicted. Some models use all available predictor variables, whether they are highly informative or not, while others apply some form of variable selection to remove uninformative variables from the model. Depending on the algorithm used to build the predictive model, it may be useful to examine the correlation between variables and remove highly correlated variables (multicollinearity problem in regression analyses). The impact of missing data is largely related to the choice of learning algorithm. Missing values in a data set can be handled in several ways, and the approach used depends on whether the data are missing because they are unknown or because they are not applicable. Some algorithms can make predictions even when some values are unknown, missing values are simply not used in the prediction.
Model building	Machine Learning algorithms are used to build the model.
Model evaluation	In order to evaluate the quality of the predictive model, a test data set with known labels is required. The predictions made by the model on the test set can be compared to the known true labels of the test set to evaluate the model. A wide variety of measures are available for comparing the similarity of known true labels and predicted labels, such as accuracy, precision, and recall.

2.4 Machine Learning algorithms

Machine Learning (ML) is one of the most active and exciting areas of computer science today. It is a branch of artificial intelligence that deals with designing algorithms that improve their efficiency based on empirical data. Machine Learning algorithms learn information and relationships between them directly from data.

Machine Learning and Data Mining often use the same methods and overlap significantly, but while Machine Learning focuses on prediction (based on known properties learned from training data) Data Mining focuses on discovering (previously) unknown properties in data.

The three main areas of Machine Learning are: supervised, unsupervised and reinforcement learning (Figure 5).

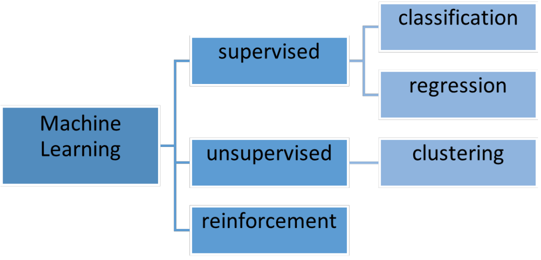

Figure 5. Areas of Machine Learning.

Supervised learning predicts the values of output variables based on input data. The model is developed from training data in which the values of the input and output variables are defined. The model generalizes the relationship between input and output variables and uses them to predict other data sets where only the input data is known.

The two main models of supervised methods are classification and regression. In classification, the output variable is discretely valued, and in the regression problem, the output variable is continuously valued. Classification algorithms determine which of the predefined categories the input data belongs to. The task of regression algorithms is to predict the numerical value of the output variable after specifying the input variable.

Below are briefly described the most common Machine Learning classification and regression algorithms used to build predictive models in EDM and LA domains.

- Linear regression (LR): The simplest form of regression is a simple linear regression that tries to fit the data set to a straight line. This is possible if the relationship between the input and output variables is linear. In multiple linear regression, there are multiple independent variables (inputs/predictors) and one dependent variable (output/response/target).
- Logistic Regression (LogR): Used to solve classification problems. While linear regression deals with predicting the value of a continuous output variable, logistic regression deals with predicting a categorical output variable. Based on the values of the input variables, it returns a binomial result (probability whether an event occurs or not, in the form of 0 and 1) or a multinomial result (several predefined possible outcomes).

- K-Nearest Neighbors (kNN): Uses a database where data points are grouped into several classes, and the algorithm tries to classify the data sample given as a classification problem.

- Decision Tree (DT): Used to solve classification and regression problems. Separates data based on a specific parameter. Data is divided into nodes and decisions are in leaves. In the classification tree, the decision variable is categorical (result in the form of yes/no), and in the regression tree, the decision variable is continuous.

- Naive Bayes Algorithm (NB): One of the simpler classification algorithms. It is a probabilistic classifier based on Bayes' theorem. The basic assumption is the mutual independence of the input variables and is therefore called naive. In real problems, the assumption that all input variables are independent of one another can rarely apply. There are three basic types of this classifier: Gaussian, Multinomial, and Bernoulli.

- Support Vector Machines (SVM): Solves classification and regression problems. In this method it is necessary to define the hyperplane which is the decision boundary. If there is a set of objects belonging to different classes, then a decision level is needed to separate them. Objects may or may not be linearly separable. If not, complex mathematical functions called kernels are needed to separate objects that are members of different classes.

- Artificial Neural Network (ANN): Is a set of interconnected simple process elements, units or nodes whose functionality is based on a biological neuron (Figure 6). At the same time, the computing power of the network is stored in the strength of the connections between individual neurons, the weights that are achieved through the adaptation process, i.e. through learning from the learning data set. A neural network processes data through distributed parallel work of its nodes. Deep learning is a subfield of ANNs, so named because it uses multi-layer neural networks to process data.

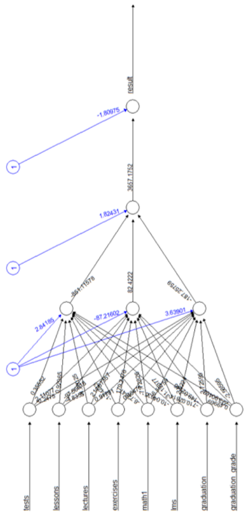

Figure 6. Graphic representation of ANN.

Ensembles are often used in the development of Machine Learning algorithms. They combine multiple individual models to create a more efficient predictive model, improving the accuracy of classification and regression models. Examples of ensembles are:

- Stacking: Involves training a learning algorithm to combine the predictions of several other learning algorithms. First, all other algorithms are trained using the available data, and then the combinator algorithm is trained to produce a final prediction using all the predictions of the other algorithms as additional inputs.

- Boosting: Involves sequential addition of set members that correct predictions made by previous models and yields a weighted average of predictions.

- Bagging: Builds decision trees in parallel, and at the end takes the most frequent prediction value from those decision trees and arrives at the right prediction. The bagging is also used in Random Forest (RF). Random Forest builds several Decision Trees and then "merges" them in order to obtain the best and most stable predictions. The so-called merging is actually done by a bagging algorithm.

3 Overview of previous research

This section reviews the relevant literature on Learning Analytics and Educational Data Mining over the last five years.

A comprehensive review on Educational Data Mining and Learning Analytics in higher education, published in 2019 entitled "Educational Data Mining and Learning Analytics for 21st Century Higher Education: A Review and synthesis" was published by the authors Aldowah, Al-Samarraie and Fauzy [23]. It is the most cited paper in the last five years in the Web of Science CC database for search TI=(Educational Data Mining) AND TI=(Learning Analytics). This review covered the most relevant work from 2000 to 2017 related to four main areas of computer-based analytics: learning (computer-based learning analytics, CSLA), predictive (computer-based predictive analytics, CSPA), behavioral (computer-based behavior analytics, CSBA), and visualization (computer-based visualization analytics, CSVA). Based on an analysis of 402 articles, it was determined that specific EDM and LA techniques can provide the best solutions to specific learning problems related to CSLA, CSPA, CABA and CSVA. Key Data Mining techniques such as clustering, association rule, visual data mining, statistics, and regression proved suitable for use in the LA and EDM domains. However, this review found that some techniques, such as sequential pattern mining, text mining, correlation mining, outlier detection, causal mining and density estimation, need not be used frequently due to the complexity of obtaining the attributes needed.

The research "Educational Data Mining and Learning Analytics: An updated survey" by authors Romero and Ventura [24] from 2020 is an complement and improved version of the previous research published in 2013. It is the most cited work in the last five years in the Scopus database for search TITLE (Educational Data Mining) AND TITLE (Learning Analytics). This paper provides an overview of major publications, major milestones, knowledge discovery cycles, major educational environments, specific tools, freely available datasets, commonly used methods, major goals, and future trends in this research area.

A systematic literature review entitled “Educational Data Mining for Student Performance Prediction” [25] aimed to identify emerging trends and methods for predicting student academic performance in research from 2015 to 2021. The authors review 58 research articles from the Lens and Scopus databases and show that the research focus of the articles is on identifying factors that affect student performance, the performance of data mining algorithms, and data mining related to e-learning systems. The authors also state that academic and demographics factors are the most important factors influencing academic success. The most commonly used approach is classification, and the Decision Tree classifier is the most commonly used algorithm.

Authors of the review “Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review” [26] indicate that their research is among the first to synthesize intelligent models and paradigms used in education and apply them to predict student learning outcomes, which are an indicator of student performance. They analyzed a total of 62 relevant papers that focused on three perspectives; forms in which learning outcomes are predicted, predictive analytic models developed to predict student learning outcomes, and dominant factors that influence learning outcomes. They concluded that achievement of learning outcomes was mainly measured as within-group scores (i.e. ranks) and as performance scores (i.e. grades). Regression and supervised Machine Learning models have been widely used to classify student performance. The best predictors of learning outcomes were students’ online learning activities, intermediate grades, and students’ academic emotions.

The main objective of the study "Factors Affecting Students’ Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques" [27] was to identify the most commonly studied factors affecting student performance, and the most common data mining techniques used to identify these factors. The results showed that the most common factors were grouped into four main categories, namely students’ previous grades and class performance, students’ e-learning activities, students’ demographic data, and students’ social information. In addition, the results showed that the most commonly used data mining techniques to predict and classify the students’ factors are Decision Drees, Naïve Bayes classifiers, and Artificial Neural Networks.

The study "Predicting academic success in higher education: literature review and best practices" [28] provides educators with easier access to Data Mining techniques to realize the full potential of their application in education. The two most important factors for predicting student success, namely prior academic performance and student demographic characteristics, were reported in 69% of the research papers. As for the prediction techniques, many algorithms were applied to predict student success in the classification technique; Decision Tree algorithms (44%), Bayesian algorithms (19%), Artificial Neural Networks (10%), Rule learner’s algorithms (9%), Ensemble Learning (7%), k-Nearest Neighbor (5%). WEKA was the most commonly used predictive modeling tool. The most commonly used measures in the literature are Accuracy, Recall, Precision, Specificity, F-Measure and ROC-curve.

Authors of the paper "Predicting students performance using educational data mining and learning analytics: A systematic literature review" [29] aims to conduct a systematic literature review of predicting student’s performance using educational data mining and learning analytics to identify techniques, attributes, metrics used, and to define the current state of the art. The most commonly used prediction techniques were Decision Tree, Regression, and Neural Network. It is also observed that most studies tend to predict course performance (pass, fail), course scores/marks, student at-risk of dropout/attrition/retention in traditional/online/blended context. Course performance, log data, midterm marks/assignments/quiz marks, demographics were the most commonly used attributes used to predict the performance.

The paper "Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022)" [30] provides a systematic review of recent studies in PLA to identify current trends and advances in the field. The results show that most of the existing publications use predictive models to assess student performance and predict those at risk of failing or dropping out. In terms of the techniques used for prediction, Artificial Neural Networks, Random Forest, and Gradient Boosting placed first, second, and third, respectively, in terms of prediction accuracy and usage frequency in comparison to other algorithms. The performance of the algorithms was commonly evaluated using the confusion matrix and the measurements obtained from it.

Through own systematic literature review, described in the next section of this paper, we aim to find and analyze relevant work on various predictions in higher education using Machine Learning algorithms, in order to contribute to research in the areas of Learning Analytics and Educational Data Mining.

4 Methodology

The purpose of this systematic literature review is to examine the use of predictive methods in higher education based on Machine Learning algorithms in the fields of Learning Analytics and Educational Data Mining.

The purpose is achieved by answering the following research questions:

Q1: How was the data collected for the research, what type of data was collected, and how much data was used for the research?

Q2: For what purpose is the prediction used and what is the type of the target variable?

Q3: How many different Machine Learning algorithms are used in a single study and what algorithms are used?

Q4: What evaluation measures are used for prediction?

Q5: What additional Predictive Modeling techniques do the authors cite?

Q6: Do the authors indicate what environment (software or programming language) they used for prediction?

It is important to examine how much data was used in the study, considering the number of students and the number of attributes observed, what predictive attributes are used, and how the data was collected (Q1). We want to identify the purpose of the prediction and the type of variables being predicted (Q2). Whether regression or classification Machine Learning algorithms were used depends on the nature of the target variables, and we will analyze how many different algorithms were used in a study and which algorithms were used (Q3). Based on (Q4), we want to investigate what measures are used to measure efficiency and whether authors report the most efficient algorithm. It is also important to investigate whether the authors indicate additional Predictive Modeling methods they used in addition to Machine Learning algorithms, e.g., variable selection, filling missing values, resampling methods, etc. (Q5). There are several software supports and programming languages for Machine Learning, and based on (Q6), we want to investigate whether the authors indicate the use of any of these environments.

The survey was conducted in December 2022 in two databases, Web of Science CC and Scopus. We searched for articles from English language journals and proceedings that are open access and published between 2018 and 2022.

We searched for the keywords "learning analytics" or "educational data mining" in the title only and "predict*" in the title only and "machine learning" and "student*" in the title and/or abstract and/or keywords.

The search in the Web of Science database CC was TI=(educational data mining) OR TI=(learning analytics) AND TI=(predict*) AND AB=(machine learning) AND AB=(student*) AND PY=(2018 OR 2019 OR 2020 OR 2021 OR 2022) AND LA=(English) NOT DT=(Review article) AND DT=(Article OR Proceeding Paper) and Open Access.

The inquiry at the Scopus database was TITLE (learning AND analytics) OR TITLE (educational AND data AND mining) AND TITLE(predict*) AND TITLE-ABS-KEY (student*) AND TITLE-ABS-KEY (machine AND learning) AND LIMIT-TO(OA , "all") AND LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO(DOCTYPE, "cp") OR LIMIT-TO(DOCTYPE, "English") OR LIMIT-TO(DOCTYPE, "2022") OR LIMIT-TO (DOCTYPE, "2021") OR LIMIT-TO(DOCTYPE, "2020") OR LIMIT-TO(DOCTYPE, "2019").

As a result of the search, 9 papers were found in the Web of Science database and 21 papers were found in the Scopus database. A total of 30 papers were found in these databases, 9 of which were included in both databases, leaving 21 papers for analysis.

Due to the relatively small number of analysis papers, the search was expanded with a new query for the keywords "machine learning" and "predict*" and "student*" in the title only and "learning analytics" or "educational data mining" in the title and/or abstract and/or keywords.

A new query in the Web of Science database was TI=(machine learning predict* student*) NOT TI=(review) AND AB=(educational data mining) OR AB=(learning analytics) AND PY=(2018 OR 2019 OR 2020 OR 2021 OR 2022) AND LA=(English) NOT DT=(Review article) AND DT=(Article OR Proceeding Paper) and Open Access.

A new inquiry at the Scopus database was TITLE(machine AND learning) AND TITLE(predict*) AND TITLE(student*) AND TITLE-ABS-KEY(educational AND data AND mining) OR TITLE-ABS-KEY (learning AND analytics) AND NOT TITLE (review)) AND (LIMIT-TO (OA, "all“) AND (LIMIT-TO(PUBYEAR, 2022) OR LIMIT-TO (PUBYEAR, 2021) OR LIMIT-TO(PUBYEAR, 2020) OR LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018)) AND (LIMIT-TO (DOCTYPE, "cp") OR LIMIT-TO (DOCTYPE, "ar") AND LIMIT-TO (LANGUAGE, "English").

As a result of the search, 8 papers were found in the Web of Science and 30 papers were found in the Scopus database. A total of 38 papers were found in these databases, of which 7 papers were included in both databases, leaving 31 papers for analysis.

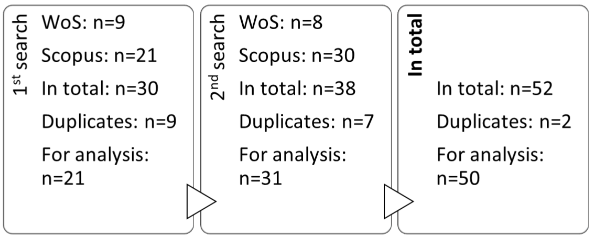

Figure 7. Search results of Web of Science CC and Scopus databases.

By combining the results of both searches, 52 papers were found, with two papers appearing in both searches, leaving 50 papers for analysis (Figure 7).

After reading the abstracts, the papers that were unimportant for this study were sorted out, 25 papers were sorted out according to different criteria, and finally 25 papers remained for detailed analysis (Table 4).

Table 4. Number of papers excluded after reading the abstracts.

Criterion	**Number of papers that do not meet the criterion**	**Reason**
The paper is written in English.	2	Only the abstract is in English.
The work is devoted to higher education.	10	6 papers deal with predictions in secondary education and 4 work with predictions in MOOCs.
Machine Learning algorithms are used in research.	2	Work does not use Machine Learning algorithms.
Work is in open access.	2	Work is not open access.
The paper is not a review paper or a meta-analysis.	6	Four review papers and two meta-analysis were excluded.
The prediction is tied to students.	3	The prediction is tied to professors.
Total	25	

Figure 8 shows the distribution of papers by country where the research was conducted, while Figure 9 shows the distribution by year when the papers were published.

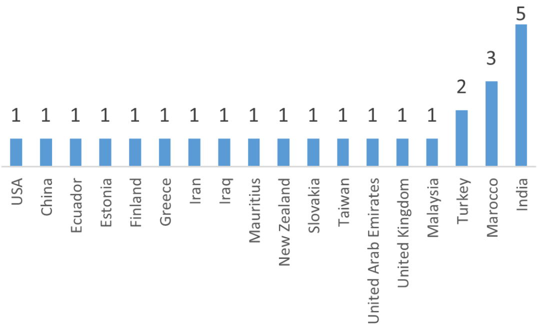

Figure 8. Distribution of analyzed papers by country in which the research was conducted.

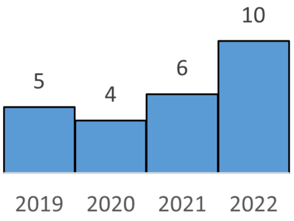

Figure 9. Distribution of analyzed papers by year of publication.

5 Findings

This section discusses the collected results of the literature review regarding the defined research questions. In order to answer the research questions, a table was created to record important facts during the reading and analysis of the works (Appendix A).

Q1: *How was the data collected for the research, what type of data was collected, and how much data was used for the research?*

The source of data collection as well as the type and amount of data collected was observed.

There are three sources of data collection: a student database, a learning management system (LMS), and a questionnaire. Data from the faculty database were used in 14 papers, data from LMS were used in three papers, and data collected by questionnaire were used in two papers. Two different sources of data collection were mentioned in six papers, with four papers mentioning the faculty database and LMS, and two papers mentioning the faculty database and the questionnaire.

Four types of input data are used for predictions in higher education: demographic data (e.g., gender and age), pre-enrollment data (e.g., grades in previous educational level), post-enrollment data (e.g., points or grades in a particular course or activity), and click-based data (e.g., time spent on a particular e-activity and participation in discussion forums). Eight papers report the use of only one type of input data, with five papers reporting the use of post-enrollment data, two papers reporting the use of click-based data, and one paper reporting the use of pre-enrollment data. Authors of 11 papers report using two types of input data, of which six papers use demographic and pre-enrollment data, three papers use demographic and post-enrollment data, and two papers use click-based and post-enrollment data. Three papers use three types of input data (demographic, pre-enrollment, and post-enrollment) and three papers use all types of input data.

Of 25 analyzed papers, four papers do not specify the amount of data used, i.e., the number of instances (students) and the number of characteristics (attributes or features). In three papers, multiple data sets were used, and for these papers, the set with the largest number of data is listed. The distribution of the number of students whose data were used for the research is shown in Figure 10. The highest number of students is 11001 and the lowest number of students is 69. The distribution of the number of features is shown in Figure 11. The lowest number of features is four, and the highest number of features is 68. It should be noted that this is only the initial number of features and that the non-informative features were later removed and only four informative features of the initial 68 were included in the predictive model.

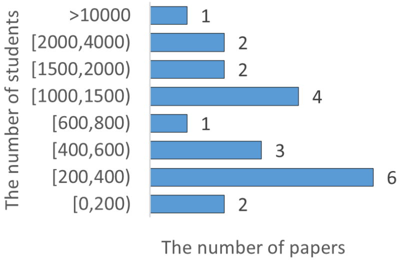

Figure 10. Distribution of analyzed papers by the number of students whose data were used for research.

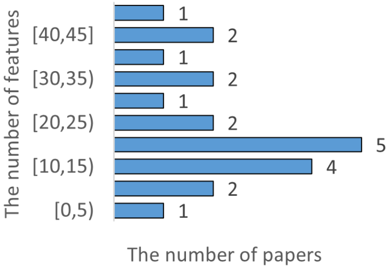

Figure 11. Distribution of analyzed papers by initial number of attributes.

Q2: For what purpose is the prediction used and what is the type of the target variable?

There are two types of variables: qualitative or categorical and quantitative or numerical. In one paper both types of target variables are observed, in two papers the target variable is numerical, while in the remaining 22 papers the target variable is categorical. Categorical variables are distinguished between nominal and ordinal variables. A nominal variable that has only two unordered categories (pass/reject, 0/1, success/fail) is called a dichotomous variable, and such target variables occur in the largest number of papers, 13. Ordinal variables with ordered categories occur as target variables in seven papers, and both dichotomous and ordinal target variables are observed in two papers. Only one paper does not specify the type of categorical target variable.

Most studies, namely eight, aim to predict at-risk students or students at risk of failing the course. The target variable in six papers analyzed is the students' final grade, which is an ordinal variable with ordered categories. Predicting an appropriate academic program or field of study is the outcome variable in three papers. Whether students drop out is the outcome variable in two studies. One study predicts the number of students enrolled, whether the admitted student will enroll in faculty, the likelihood that the student will be admitted, and the student's behavior (whether the student delays fulfilling his or her obligations). One study has two different outcome variables (graduation rate and commitment level) and one study has three different outcome variables (graduation rate, at-risk students, and dropout students).

Q3: How many different Machine Learning algorithms are used in a single study and what algorithms are used?

Regression algorithms are used in two papers, classification in 22 papers, and classification and regression algorithms in one paper. The total number of algorithms used is 126, and the distribution of the number of algorithms used in a single study is shown in Figure 12.

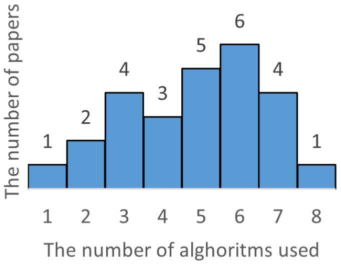

Figure 12. Distribution of analyzed papers according to the number of Machine Learning algorithms used.

In 23 studies, the predictive task was of classification type, and the total number of algorithms used in these studies is 115. Figure 13 shows the distribution of classification algorithms used.

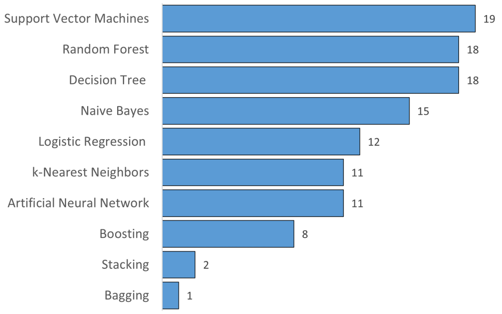

Figure 13. Distribution of classification algorithms used in the analyzed papers.

In three studies, the prediction task was of regression type and the total number of algorithms used in these studies was 11. Table 5 shows the regression algorithms used in these three studies.

Table 5. Presentation of regression algorithms used.

Algorithm	Number of algorithms used in research
Linear Regression Support Vector Regression Decision Tree Regression Random Forest Regression	4
Boosted Tree Regression Tree	2
Random Forest M5Rules Bagging SMOreg IBk-5NN	5

Q4: What evaluation measures are used for prediction?

When evaluating a Machine Learning algorithm, its performance is measured on an unused data set, a test set. Efficiency is measured with a measure or set of measures specific to a prediction task (classification or regression). The purpose of algorithm evaluation is to assess the performance of the algorithm on test data and compare it to other algorithms or models.

Different predictive tasks and data may require different evaluation measures. In some cases, more than one evaluation measure is used to better understand the performance of the model.

Table 6 lists the evaluation measures used in the classification algorithms of the analyzed papers, their brief description, and the frequency of use of a particular measure.

Table 6. An overview of the evaluation measures used for classification.

Evaluation measure	**Description**	**Number**
Accuracy	proportion of correctly classified instances in a dataset	23
Precision	proportion of accurately classified in a set of positively classified examples	18
Recall (Sensitivity)	proportion of accurately classified in the set of all positive examples	18
F1	the harmonic mean of precision and sensitivity, balances both measures	14
AUC-ROC	area below the ROC curve, compares the true positive rate with the false positive rate	12
TP	number of accurate positive predictions	3
FP	number of incorrect positive predictions	3
Kappa index	shows the degree of agreement between the frequencies of two data sets collected on two different occasions	2

A summary of the evaluation measures used for regression algorithms, their brief description, and the number of papers in which each measure has been used can be found in Table 7.

Based on the calculated evaluation measures, the authors of the analyzed papers indicate which of the algorithms used is the most effective.

In 15 studies the target variable was dichotomous. In these papers the authors cite four times LogR, two times RF, two times Stacking and once DT, SVM, NB, Boosting and Auto- WEKA as the most effective algorithm. The authors of two papers mention the two most efficient algorithms, first RF and ANN, second LogR and SVM. In 11 papers, when stating the most efficient algorithm, its accuracy is listed, in one paper AUC-ROC and in three papers all calculated evaluation measures.

Table 7. An overview of the evaluation measures used for regression.

Evaluation measure	**Description**	**Number**
R^2	the proportion of variance in the target variable explained by the predictions of the model	1
MEAN ABSOLUTE ERROR (MAE)	the average absolute difference between the predicted and the true values	2
MEAN SQUARE ERROR (MSE)	the share of accurately classified examples in the set of all positive examples	1
RMSE	the second root of the mean square error	2

In nine studies the target variable was ordinal, and in these studies the authors cite RF four times and LogR on time as the most efficient algorithm. The authors of four papers list multiple most efficient algorithms, two times RF and ANN and once ANN and SVM, ANN and SVM and RF and DT. In eight papers, when listing the most efficient algorithm, its accuracy is given and in one paper the measure F1.

In only three studies the target variable was numerical, and due to the small number of papers it is not possible to define the most effective algorithm.

Q5: What additional Predictive Modeling techniques do the authors cite?

Data preprocessing is the process of cleaning, transforming, and organizing a data set before it is input into a Machine Learning model. This step is critical to the performance of the model, as it can help improve the quality of the data and make it more suitable for a particular prediction task. The preprocessing techniques used in the analyzed papers are listed in Table 8.

Table 8. Presentation of preprocessing techniques used in the analyzed papers.

Technique	**Number of papers**
Selection of informative variables or removal of non-informative variables.	19
Data transformation or conversion of numerical variables into categorical.	9
Normalizing or standardizing data.	9
Resampling a dataset to reduce the imbalance ratio.	6
Testing the correlation of variables.	6
Replacing missing values.	5
Delete missing values.	4
Handling outliers.	2

Validation in Machine Learning involves evaluating the performance of a model against a dataset that is separate from the training data. The purpose of validation is to evaluate the efficiency of the model on unused data and adjust the model's hyperparameters to optimize its efficiency. It is important to use a validation method that is appropriate for the size and characteristics of the data set. The authors of the 20 papers analyzed list the use of some of the validation methods listed in Table 9.

Table 9. Presentation of validation methods used in the analyzed papers.

Validation method	Number of papers
Holdout method: The data is divided into a training set and a testing set, the model is trained on a training set and evaluated on a testing set.	4
K-fold validation: The dataset is divided into K subsets, the model is trained on K-1 subsets and evaluated on the remaining subsets. This process is repeated K times, each subset once serves as a test set.	1
Leave-one-out: Variant of K-fold validation, but with K = n (n is the number of examples in the dataset)	2
K-fold cross-validation: A combination of K-fold and Leave-one-out validation, the model is trained and evaluated multiple times using different data partitions.	12
Bootstrap validation: The model is trained and evaluated several times by random sampling of data with replacement.	1

Hyperparameter optimization is the process of adjusting algorithm parameters to optimize its work on a given task. Common techniques for hyperparameter optimization include network search, random search, and Bayesian optimization. Hyperparameter optimization is mentioned in five of the 25 papers analyzed.

Variable importance in Machine Learning refers to the technique of determining the relative importance of the input variables used in the model. In this way, the variables that have the greatest impact on the model prediction are determined. In the analyzed papers, this technique is mentioned four times.

Q6: *Do the authors indicate what environment (software or programming language) they used for prediction?*

The authors of eight papers indicate the use of the Python programming language, in seven papers the use of the software WEKA, and in one paper the use of the R programming language and the software WEKA. In two papers the use of the tool Orange is listed, and in one paper the use of the tools Hadoop and XLSTAT. The authors of one paper list the use of several tools, IBM SPSS, R, KNIME and Bayesian Labs. One paper uses a custom web application, while the authors of three papers do not specify the use of an environment.

6 Discussion

With respect to the six defined research questions, we examined what is predicted in higher education, what input data were used, how many Machine Learning algorithms were used in each study, and which were most effective. We also examined what other predictive modeling techniques were listed and whether a programming environment was used for prediction.

The most common number of students whose data were used in the survey ranged from 200 to 400, and the number of predictor variables entered ranged from 15 to 20. In most studies, the data were drawn from the faculty database and included demographic data and data on previous level of study. These findings support recent researches [25], [28].

Romero and Ventura in [24] state current applications or topics of interest of EDM/LA research community. One of these is early warning systems or the prediction student performance and identify at-risk students as early as possible in order to intervene early and promote student success. Also, the results of [29] and [30] show that most of the existing publications use predictive models to assess student performance and predict those who are at risk of failing or dropping out. This is consistent with our findings - most of the studies, namely eight, are concerned with predicting at-risk students, followed by the frequency of predicting students' final grade in six papers.

The most frequently mentioned algorithms in the analyzed papers are Support Vector Machine, Random Forest, Decision Tree, Naive Bayes, and Logistic Regression. The most frequent number of algorithms used in each study is six. The use of these algorithms is also stated in [25], [27–30].

The classification problem occurs in 23 of the 25 papers analyzed. Classification is the most frequently used technique for solving predictive problems and these result is in line with the reviews [23], [25], [28]. For a dichotomous target variable, Logistic Regression is most often cited as the most successful algorithm, while for the ordinal target variable, Random Forest is cited. In testing which of the algorithms is the most successful, the authors use different evaluation measures ([28], [30]), but when they indicate the most successful algorithm, they almost always show its accuracy.

Additional predictive modeling techniques such as preprocessing, validation, and hyperparameter optimization, as well as the frequency of use of each technique, are also listed.

Finally, the most common environments in which the authors made predictions are WEKA and Python. These results support the report [28] which states that WEKA was the most commonly used tool for predictive modeling.

7 Conclusion and recommendations for future research

This paper describes a systematic literature review of predictive methods in Learning Analytics and Educational Data Mining in higher education based on Machine Learning algorithms. In selecting the papers for analysis, the main criterion was to find papers that use Machine Learning algorithms for prediction in higher education in the areas of Learning Analytics and Educational Data Mining.

Twenty-five papers were included in the detailed analysis, and six research questions were posed. We examined what is being predicted in higher education, what input data was used, how many Machine Learning algorithms were used in each study, and which were the most effective. In addition, we examined what other predictive modeling techniques were listed and whether a programming environment was used for prediction.

It should be emphasized that there are no general rules or procedures for predictions in education, including higher education. The approach depends mainly on the target variable and the input data.

The ultimate goal of any Learning Analytic process is to take pedagogical action to improve the learning and teaching process. The work analyzed does not indicate whether pedagogical actions were taken, nor does it define the theoretical framework.

Finally, this area needs further research to provide a broader pedagogical-technological framework to help teachers in the higher education system to build a predictive model and improve the learning and teaching process through appropriate pedagogical interventions.

References

[1] N. Hoić-Božić and M. Holenko Dlab, *Uvod u e-učenje: obrazovni izazovi digitalnog doba*. Rijeka: University of Rijeka, Faculty of Informatics and Digital Technologies, 2021.

[2] C. Romero and S. Ventura, “Data mining in education,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 12–27, 2013.

[3] G. Siemens and R. Baker, “Learning analytics and educational data mining: Towards communication and collaboration,” *ACM International Conference Proceeding Series*, 2012.

[4] L. Calvet Liñán and A. Juan Pérez, “Educational data mining and learning analytics: differences, similarities, and time evolution,” *RUSC. Universities and Knowledge Society Journal*, vol. 3, no. 12, pp. 98–112, 2015.

[5] A. Azevedo, “Data mining and knowledge discovery in databases,” in *Encyclopedia of Information Science and Technology*, 4th ed. IGI Global, 2018, pp. 1907–1918.

[6] S. Hussain, N. Abdulaziz Dahan, F. Ba-Alwib, and R. Najoua, “Educational data mining and analysis of students’ academic performance using weka,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, pp. 447–459, 02 2018.

[7] R. S. Baker and K. Yacef, “The state of educational data mining in 2009: A review and future visions,” *Journal of Educational Data Mining*, vol. 1, no. 1, p. 3–17, Oct. 2009. [Online]. Available: https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8

[8] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and techniques*, 3rd ed. Elsevier, 2011.

[9] C. Romero and S. Ventura, “Educational data mining: A survey from 1995 to 2005,” *ScienceDirect*, vol. 33, no. 1, pp. 135–146, 2007.

[10] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, “Analyzing undergraduate students’ performance using educational data mining,” *Computers Education*, vol. 113, pp. 177–194, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360131517301124

[11] D. Gasevic, A. Wolff, C. Rose, Z. Zdrahal, and G. Siemens, “Learning analytics and machine learning,” in *LAK 2014*, S. Teasley and A. Pardo, Eds. United States of America: Association for Computing Machinery (ACM), 2014, pp. 287–288, international Learning Analytics amp; Knowledge Conference 2014, LAK 2014 ; Conference date: 24-03-2014 Through 28-03-2014. [Online]. Available: https://lak14indy.wordpress.com/

[12] G. Kostopoulos, S. Kotsiantis, C. Pierrakeas, G. Koutsonikos, and G. Gravvanis, “Forecasting students’ success in an open university,” *International Journal of Learning Technology*, vol. 13, p. 26, 01 2018.

[13] C. Angeli, S. K. Howard, J. Ma, J. Yang, and P. A. Kirschner, “Data mining in educational technology classroom research: Can it make a contribution?” *Computers Education*, vol. 113, pp. 226–242, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360131517301264

[14] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, “Predicting student performance from lms data: A comparison of 17 blended courses using moodle lms,” *IEEE Transactions on Learning Technologies*, vol. PP, pp. 1–1, 10 2016.

[15] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. D. Kloos, “Prediction in moocs: A review and future research directions,” *IEEE Transactions on Learning Technologies*, vol. 12, no. 3, pp. 384–401, 2019.

[16] C. Márquez, A. Cano, C. Romero, and S. Ventura, “Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data,” *Applied Intelligence*, vol. 38, pp. 315–330, 08 2013.

[17] P. Long and G. Siemens, “Penetrating the fog: Analytics in learning and education,” *Educause Review*, vol. 46, no. 5, pp. 30–40, 2011.

[18] SNOLA, “Learning analytics 2018 – an updated perspective,” accessed 25/01/2023.

[Online]. Available: snola.es/2018/02/21/learning-analytics-2018-updated-perspective

[19] C. Renata Ivanković, "Learning analytics and educational data mining," accessed 25/01/2023. [Online]. Available: https://radovi2015.cuc.carnet.hr/modules/request.php?module=oc_program&action=view.php&a=&id=87&type=5

[20] SOLAR, "What is learning analytics?" accessed 25/01/2023. [Online]. Available: solaresearch.org/about/what-is-learning-analytics

[21] Moodle, "Analytics." [Online]. Available: https://docs.moodle.org/400/en/Analytics

[22] C. Brooks and C. Thompson, "Predictive modelling in teaching and learning," in *Handbook of Learning Analytics*, 2nd ed., C. Lang, G. Siemens, A. Friend Wise, D. Gašević, and A. Merceron, Eds. SoLAR, 2022, pp. 29–37.

[23] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics and Informatics*, vol. 37, pp. 13–49, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736585318304234

[24] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 05 2020.

[25] M. H. B. Roslan and C. J. Chen, "Educational data mining for student performance prediction: A systematic literature review (2015-2021)," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 17, pp. 147–179, 3 2022.

[26] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Applied Sciences*, vol. 11, p. 237, 12 2020.

[27] A. Abu Saa, M. Al-Emran, and K. Shaalan, "Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques," *Technology, Knowledge and Learning*, vol. 24, no. 4, p. 567 – 598, 2019.

[28] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, 2020.

[29] A. Dhankhar, K. Solanki, S. Dalal, and Omdev, "Predicting students performance using educational data mining and learning analytics: A systematic literature review," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 59, p. 127 – 140, 2021.

[30] N. Sghir, A. Adadi, and M. Lahmer, "Recent advances in predictive learning analytics: A decade systematic review (2012–2022)," *Education and Information Technologies*, 2022.

[31] D. Aggarwal, S. Mittal, and V. Bali, "Prediction model for classifying students based on performance using machine learning techniques," *International Journal of Recent Technology and Engineering*, vol. 8, pp. 496–503, 2019.

[32] Y. Badal and R. Sungkur, "Predictive modelling and analytics of students' grades using machine learning algorithms," *Education and Information Technologies*, 2022.

[33] K. Basu, T. Basu, R. Buckmire, and N. Lal, "Predictive models of student college commitment decisions using machine learning," *Data*, vol. 4, 2019.

[34] A. Bayazit, N. Apaydin, and I. Gonullu, "Predicting at-risk students in an online flipped anatomy course using learning analytics," *Education Sciences*, vol. 12, 2022.

[35] D. Buenaño-Fernández, D. Gil, and S. Luján-Mora, "Application of machine learning in predicting performance for computer engineering students: A case study," *Sustainability (Switzerland)*, vol. 11, 2019.

[36] S. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, H. Fujita, and N. Ghani, "Multiclass prediction model for student grade prediction using machine learning," *IEEE Access*, vol. 9, pp. 95 608–95 621, 2021.

[37] A. Dirin and C. Saballe, "Machine learning models to predict students' study path selection," *International Journal of Interactive Mobile Technologies*, vol. 16, pp. 158–183, 2022.

[38] R. Ghorbani and R. Ghousi, "Comparing different resampling methods in predicting students' performance using machine learning techniques," *IEEE Access*, vol. 8, pp. 67 899–67 911, 2020.

[39] C. C. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," *Computers Education*, vol. 131, pp. 22–32, 4 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360131518303191

[40] I. Guabassi, Z. Bousalem, R. Marah, and A. Qazdar, "A recommender system for predicting students' admission to a graduate program using machine learning algorithms," *International journal of online and biomedical engineering*, vol. 17, pp. 135–147, 2021.

[41] A. Hashim, W. Awadh, and A. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," vol. 928, 2020.

[42] D. Hooshyar, M. Pedaste, and Y. Yang, "Mining educational data to predict students' performance through procrastination behavior," *Entropy*, vol. 22, p. 12, 2020.

[43] T.-T. Huynh-Cam, L.-S. Chen, and K.-V. Huynh, "Learning performance of international students and students with disabilities: Early prediction and feature selection through educational data mining," *Big Data and Cognitive Computing*, vol. 6, 2022.

[44] J. Jayapradha, K. Kumar, and B. Deka, "Educational data classification and prediction using data mining algorithms," *International Journal of Recent*

Technology and Engineering, vol. 8, pp. 8674–8678, 2019.

[45] D. Jeslet, D. Komarasamy, and J. Hermina, "Student result prediction in covid-19 lockdown using machine learning techniques," vol. 1911, 2021.

[46] J. Kabathova and M. Drlik, "Towards predicting student's dropout in university courses using different machine learning techniques," *Applied Sciences (Switzerland)*, vol. 11, 2021.

[47] T. Kumar, K. Sankaran, M. Ritonga, S. Asif, C. S. Kumar, S. Mohammad, S. Sengan, and E. Asenso, "Fuzzy logic and machine learning-enabled recommendation system to predict suitable academic program for students," *Mathematical Problems in Engineering*, vol. 2022, 2022.

[48] N. Lebkiri, M. Daoudi, Z. Abidli, J. Elturk, A. Soulaymani, Y. Khatori, Y. E. Madhi, and M. Benattou, "Using machine learning for prediction students failure in morocco: an application of the crispdm methodology," *International Journal of Education and Information Technologies*, vol. 15, pp. 344–352, 10 2021.

[49] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Computers and Education: Artificial Intelligence*, vol. 3, 2022.

[50] F. Ouatik, M. Erritali, F. Ouatik, and M. Jourhmane, "Predicting student success using big data and machine learning algorithms," *International Journal of Emerging Technologies in Learning*, vol. 17, pp. 236–251, 2022.

[51] G. Ramaswami, T. Susnjak, and A. Mathrani, "On developing generic models for predicting student outcomes in educational data mining," *Big Data and Cognitive Computing*, vol. 6, 2022.

[52] S. Shilbayeh and A. Abonamah, "Predicting student enrolments and attrition patterns in higher educational institutions using machine learning," *International Arab Journal of Information Technology*, vol. 18, pp. 562–567, 2021.

[53] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, "Implementing automl in educational data mining for prediction tasks," *APPLIED SCIENCES-BASEL*, vol. 10, 1 2020.

[54] S. Verma, R. Yadav, and K. Kholiya, "A scalable machine learning-based ensemble approach to enhance the prediction accuracy for identifying students at-risk," *International Journal of Advanced Computer Science and Applications*, vol. 13, pp. 185–192, 2022.

[55] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, 2022.

Appendix A

	Q1			Q2		Q3			Q4		Q5	Q6
ID	**DCP**	**Types of features**	**Number of data**	**Prediction**	**Type of target variable**	**Type C/R**	**Algorithms compared**	**Number of alg.**	**Evaluation Measures**	**Outperf.**	**Methods**	**Environment**
ID01 [31]	SDB	DEM, PreE	131x22 (6)	reappear/s in the course	Dichotomous: Y/N	C	NB, LogR, MLP SVM,J48,RF	6	A, TP, FP, P, R, F1, AUC-ROC, MAE	MLP and RF (A 92.3%)	feature selection	WEKA
ID02 [32]	SDB, LMS	DEM,PreE, PostE,click	1074x42 (30)	final grade eng. level	Ordinal[5]:A,B,C,D,E Ordinal[4]:H,L,M,N	C	LogR, RF, ANN kNN,SVM,DT,NB	7	A, P, R, F1	RF(A 85% and 83%)	imput,select, res, norm, cross-val(10)	own Web application
ID03 [33]	SDB	DEM,PreE	11001x35 (15)	if applicant accepts offer	Dichotomous:0,1	C	LogR, NB, DT, SVM, kNN, RF,GB	7	A,AUC-ROC	LogR (A-R 77.8%)	selec,norm,imput, cross-val(10),hyper	Python
ID04 [34]	LMS, SDB	click, PostE	69x12	at-risk students	Dichotomous:0,1	C	kNN,DT,NB, RF,SVM	5	A,P,R,F1,A-R	NB(A 71%)	cross-val(10)	Orange
ID05 [35]	SDB	PostE	335x68 (4)	final grades	Dichotomous:F/P	C	DT	1	A	DT(A 96.5%)	select,transform	WEKA
ID06 [36]	SDB	PostE	1282x10 (6)	final grades	Ordinal [5]	C	J48(DT),SVM,NB, kNN,LogR,RF	6	A,P,R, F1	RF (F1 99.5%)	select,res, cross-val(10)	WEKA
ID07 [37]	survey, SDB	DEM,PreE, PostE	/	study path	Dichotomous	C	LogR,DT,RF	3	A,A-R,Kappa	RF (A 94%)	imput,transform, norm,select,out, coll,bootstrapped	IBM SPSS, KNIME,R Bayesian labs
ID08 [38]	SDB	DEM,PreE, PostE	650x19, 394x19	final GPA	Ordinal[4], Dichotomous	C	RF,kNN,ANN,XGB, SVM, DT,LogR,NB	8	A,R,P,F1	ANN,RF LogR	res,norm, cross-val(random)	Python
ID09 [39]	SDB	PostE	4970x32 (5)	student outcomes	Ordinal[5]	C	RF,MLP,NB,DT	4	A,TP,FP,P, R, F1,A-R	RF (A 97%)	select,cross-val (leave one out,n)	WEKA
ID10 [40]	SDB	PreE	500x8	students' admission	Numeric: [0,1] probability	R	LR,SVR,DTR,RFR	4	MSE,RMSE,R^2	RFR (R^2= 89%)	corr,importance	XLSTAT
ID11 [41]	SDB	DEM,PreE, PostE	499x9	success pred., academic per.	Ordinal[6] Dichotomous	C	DT,NB,LogR,ANN, SVM,kNN,SMO	7	A,TP,FP, P,R,A-R	LogR(A 69%) LogR(A 89%)	removing, cross-val(70:30)	WEKA
ID12 [42]	LMS	click	242x16	students' performance	Ordinal[3]	C C	SVM,GP,DT,RF, ANN,ADB,NB	7	A,P,R, F1	NN(A 96%) SVM(A 95%)	cross-val(k)	/
ID13 [43]	SDB	PreE, DEM	3885x16, 104x14, 47x16	AP of local, international, stud. with dis.	Ordinal[5]	C	MLP,SVM,RF,DT	4	A,P,R,F1,A-R	RF(A 98.6%) SVM,MLP,DT SVM	select,rem,trans, norm,res,import, cross-val(80:20)	Python
ID14 [44]	survey	DEM, PreE	265x13	choice of department	Ordinal[8]	C	SVM,RF,NB	3	A,Kappa	RF (A 71%)	norm,corr,select, cross-val(k)	R, WEKA
ID15 [45]	SDB	DEM, PostE	1460x15 (3)	is st. eligible to acquire degree	Dichotomous	C	LogR,SVM	2	A,P,R	LogR,SVM (A 99.72%,P,R)	select,norm, imput,75:25	Python

	Q1			Q2		Q3			Q4		Q5	Q6
ID	**DCP**	**Types of features**	**Number of data**	**Prediction**	**Type of target variable**	**Type C/R**	**Algorithms compared**	**Number of alg.**	**Evaluation Measures**	**Outperf.**	**Methods**	**Environment**
ID16 [46]	LMS	click, PostE	261x30 (5)	student's dropout	Dichotomous	C	LogR,DT,NB, SVM,RF,ANN	6	A,P,R,F1,A-R	RF(A 93%, P,R,F1)	select,trans,corr, norm,cross-val(10),hyper	Python
ID17 [47]	SDB	DEM, PostE	1000x21 (15)	academic program	/	C	SVM,RF,DT	3	A,P,R	SVM(A)	select	WEKA
ID18 [48]	SDB, survey	DEM, PostE	/	student failure	Dichotomous	C	Stacking,DT,RF, ADB,SVM,kNN	6	A,P,R	Stacking (A 98%)	remove,transform, select, corr,70:30	/
ID19 [49]	SDB	PostE	261x12 (8)	student's dropout	Dichotomous	C	FNN,GB,RF, XGB,Stacking	5	A,P,R,F1, A-R	Stacking (A 92.18%,P,R,F1)	select,remove,transform, corr,cross-val(10),hyper	Python
ID20 [50]	SDB, LMS	DEM,PostE, PreE,click	/	student success	Dichotomous	C	kNN,SVM,DT	3	A,R,P,F1	SVM(A 87.32%)	select,transform	Hadoop
ID21 [51]	LMS, SDB	DEM,PreE, PostE,click	/	at-risk students	Dichotomous	C	CatBoost,RF,NB, LogR,kNN	5	A,R,P,F1,A-R	CatBoost (A 75%)	select, norm, res, cross-val(k),importance	Python
ID22 [52]	SDB	DEM, PreE	1600x16	student enrolments	Numerical	R	Boosted Tree, Regression Tree	2	A	BT (A 89%)	imput,outliers, cross-val(10)	/
ID23 [53]	LMS	click	282x42 180x44 129x45	student grades	Scale[0,10]	R	RF, M5Rules,Bagging, SMOreg, IBk-5NN	5	MAE	Auto-WEKA	trans,select,hyper cross-val(10),importance	WEKA
ID23 [53]	LMS	click	282x42 180x44 129x45	at-risk dropout	Dichotomous	C	NB, RF, Bagging, DT,SVM,kNN	6	A,A-R	Auto-WEKA	trans,select,hyper cross-val(10),importance	WEKA
ID24 [54]	survey	DEM, PreE	550x27 (11)	students at-risk	Dichotomous	C	DT,NB,kNN, SVM,LogR	5	A,P,R,F1	LogR (A 94.54%)	transform,select,res, val(hold-out),hyper,80:20	Python
ID25 [55]	SDB	PostE	1854x4	final exam grades	Ordinal[4]	C	RF,ANN,kNN, SVM,LogR,NB	6	A,P,R,F1,A-R	RF and ANN (A 75%)	select,cross-val(10)	Orange