The Discovery of Knowledge in Educational Databases: A Literature Review with Emphasis on Preprocessing and Postprocessing

Léo Manoel Lopes da Silva GARCIA
Faculdade de Ciências Exatas – Curso de Ciências da Computação – Universidade do Estado de Mato Grosso (UNEMAT) – Barra do Bugres – MT – Brasil
leoneto@unemat.br
ORCID: 0000-0003-4861-8830

Daiany Francisca LARA
Faculdade de Ciências Exatas – Curso de Ciências da Computação – Universidade do Estado de Mato Grosso (UNEMAT) – Barra do Bugres – MT – Brasil
dflara@unemat.br
ORCID: 0000-0002-0458-9196

Raquel Salcedo GOMES
Programa de Pós-Graduação em Informática na Educação – Universidade Federal do Rio Grande do Sul (UFRGS) – Porto Alegre – RS – Brasil
raquel.salcedo@ufrgs.br
ORCID: 0000-0001-9497-513X

Silvio Cézar CAZELLA
Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Rua Sarmento Leite, 245 - Porto Alegre - RS - Brasil
silvioc@ufcspa.edu.br
ORCID: 0000-0003-2343-893X

ABSTRACT
In educational data mining (EDM), preprocessing is an arduous and complex task and must promote an appropriate treatment of data to solve each specific educational problem. In the same way, the parameters used in the evaluation of postprocessing results are decisive in the interpretation of the results and decision-making in the future. These two steps have as much influence on obtaining good results in EDM as the algorithms used. However, in the dissemination of the results of studies on this topic, emphasis is placed only on the evaluation of the algorithms used. Thus, the present study sought to carry out a systematic review of the literature on this topic, focusing on the exploration of the preprocessing performed and on the metrics for evaluating the results. It is observed in many studies that the description and evaluation of the preprocessing and the use of several metrics to evaluate the algorithms used are negligible. However, without a proper explanation of the meaning of each metric to reach the proposed objective

INTRODUCTION
The higher education educational system comes from a long and continuous expansion of its access places, favoring social equity through education. According to the Organization for Economic Co-operation and Development (OECD), higher education drives the growth of economies by promoting the acquisition of specific skills and competences, qualifying academics for various job functions (OECD, 2019). This prerogative is foreseen in the Law of Directives and Bases of Brazilian Education, which, in article 2nd, defines the principles and purposes of national education: "Education, a duty of the family and the State, inspired by the principles of freedom and the ideals of human solidarity, aims at the full development of the student, his preparation for the exercise of citizenship and his qualification for work" (LDB, 1996). However, just providing access is not enough, as there are difficulties faced by higher education institutions that impact their efficiency in graduating professionals. The most prominent difficulties include low academic performance, evasion, and retention, which make it impossible or delay graduation (ACKCAPINAR, 2019).

Considering this, institutions and researchers have tried to understand these phenomena to support the creation of effective policies to face and resolve these problems. Thus, as in the commercial and business area, where innovative technologies have been incorporated into the management and decision-making process, in the educational area, initiatives that aim to use data analysis technologies to support educational actions in the educational process have been considered promising. In the same way, the business sector exploits the abundance of data generated continuously by systems that computerize all its processes. The educational system reveals this same potential, revealed through the vast amount of data generated by academic systems, educational platforms, exam results and activities, inside and outside virtual learning environments.

In this context, the educational data mining (EDM) approach appears, which comprises a field of research aimed at the application of data mining whose origins are inherent to the educational context, with the main intention of producing a reliable understanding of learning patterns of students and identifying their study processes and behaviors to enhance educational outcomes (ANOOPKUMAR, 2018; SOKKHEY, 2020). As a result of this deepening of information, it is believed that student performance can be improved more effectively through strategic programs (MENGASH, 2020).

In this conjecture of several possible applications and the availability of many techniques, a single EDM method that works best in all contexts and applications has not yet been identified. Often, more than one technique is explored to determine which technique leads to better results in a particular case (ALTURKI, 2021). It stands out that these techniques comprise not only the algorithms used but also the process of data collection, data processing and evaluation of results. In this sense, systematically exploring works that present studies in EDM can contribute to the identification of good practices, the recognition of varied techniques and their application potential and the signaling of gaps to be filled.

Thus, this article presents a systematic literature review (SLR) about the use of EDM in higher education in predictive approaches. Considering that there is a predominance of SLR on this topic that addresses the main algorithms used (SHAHIRI 2015, PEÑA-AYALA 2014 ALTURKI 2020) and considering that works in EDM tend to apply several algorithms in the same study, this work has the main objectives of investigating which attribute selection methods, preprocessing, and evaluation metrics are being used the most. With the assumption that these steps, in addition to the algorithms used, are fundamental to define good results in an EDM enterprise, they are reliable and potentially generalizable.

In the next section, a discussion is undertaken that seeks to compose a theoretical framework on the EDM theme, highlighting the approach carried out in some studies. Then, the proposal of the systematic review is described, along with its respective methodological details. Subsequently, the results and discussions are presented. Finally, final considerations are made, resuming and evaluating the objectives achieved.

EDUCATIONAL DATA MINING

The term “educational data mining” first appeared in 2005 during the annual conference of the Association for the Advancement of Artificial Intelligence in the city of Pittsburgh, United States (USA) (SADIQ, 2019). Since then, this technique has become popular among researchers in the technological and educational area, building a consensus that there are many benefits that EDM can bring to education (AKMEŞE, 2021; SOKKHEY, 2020; HELAL, 2018; MIGUÉIS, 2018; ASIF, 2017; SADIQ, 2019; MAGBAG, 2020; COSTA, 2017; ALTURKI, 2021; URBINA-NÁJERA, 2020; PABREJA, 2017). The purpose of educational data mining is to obtain meaningful information from data from the educational environment, with the expectation that it can provide an in-depth understanding of educational phenomena and provide substantial support for decision-making. To this end, strategies from the field of computational intelligence are employed, which involve statistical methods and data analysis, composing the data mining techniques already widely disseminated in other areas.

In data mining, approaches can be divided into two main categories: descriptive and predictive methods. Descriptive methods allow identifying patterns in the data, recognizing possible rules of cause and effect. Predictive methods aim to make inferences about the future, enabling predictions of occurrences through induction based on previous occurrences (AKMEŞE, 2021). Both approaches have numerous potential applicabilities. However, predictive methods applied to the prediction of academic performance have stood out as the most popular (MIGUÉIS, 2018; SOKKHEY, 2020; ALTURKI, 2021). Early identification of the results that students may have during or after an educational process allows teachers or managers to intervene in a timely manner and reverse an unsatisfactory academic trajectory. Alturki et al. (2021) describe three types of predictions in higher education: (i) performance prediction (average grades) at the course level; (ii) prediction of graduation or abandonment of a course, and (iii) prediction of student results in specific subjects. With the early identification of these trends, it may be possible to guard against possible malfunctions in the education process. Armed with this information, instructors can monitor students' progress and intervene early on academic problems, negotiate alternative routes and discerning strategies to fill gaps or overcome vulnerabilities (AKMEŞE, 2021).

It is important to emphasize that the use of EDM is not limited to the application of algorithms that implement mining techniques. It encompasses a rigorous process of data collection and treatment, which directly interferes with the results achieved. Sokkhey (2020) describes the EDM process in 5 steps: (i) first, it is necessary to obtain the desired data, which can be generated by academic systems, virtual learning environments, digital educational objects or collected through observation or questionnaires; (ii) the second stage comprises the processing of data, since they are hardly found “clean”, that is, and in appropriate formats, requiring preparation; (iii) in the third step,

the data are submitted to data mining algorithms, which may vary according to the objective to be achieved. In the same way, several algorithms can be used, evaluating the best performance obtained for the desired objectives; (iv) in the fourth stage, the results need to be interpreted to guide decision-making; and (v), finally, the fifth stage comprises modifying the educational processes according to the identified trends, the predictions obtained, and decisions formulated to direct them. This step can be postponed if the results are inconclusive: the process can be readjusted, and the first 4 steps can be repeated.

Several academic studies make important contributions to this area, exploring and presenting different methods of implementing each step of the EDM process. In this context, literature reviews seek to explore the methodologies used, highlighting the most commonly used procedures or those of recognized effectiveness. Shahiri et al. (2015), for example, explored which attributes are most influential in academic performance and which algorithms are most used. Their results showed that the cumulative grade average is the most commonly used attribute in EDM. This is a more common metric in American institutions, called the Grade Point Average (GPA) in English, and represents the average of your grades in each period. In Brazil, its use is not common in academic systems, but it can be easily calculated.

Peña-Ayala (2014) focused his review on mining approaches, concluding that 60% of the articles in his review used predictive methods, while 40% used descriptive methods. The author also indicated the predominance in the implementation of Bayesian network techniques, decision trees and instance-based learning. Saa et al. (2019) investigated the most influential factors in academic performance by grouping them into 4 categories, namely, students' past grades and class performance, students' e-learning activities, student demographics, and student social information. In this survey, the techniques of decision trees, naïve Bayes classifiers and artificial neural networks stood out as the most used.

It is noted that there are many aspects to be explored when carrying out a substantial analysis of one or some educational data mining processes. Although some approaches may be repeated, carrying out studies such as these will always be important due to the emergence of new works. There is a tendency to use multiple techniques and algorithms to highlight, at the end, which of them presented better results. In this sense, limiting it to only one algorithm or restricting it to a few techniques in the testing phase are not necessary when it is possible to explore several techniques until an acceptable performance is obtained. Thus, this review does not focus on examining the algorithms used, but on the preprocessing, activities carried out and how they promoted improvements or not in the results, considering that these activities are highly influential in the final performance, capable of making a data source irregular in a promising dataset, providing good results.

PROPOSAL AND METHODOLOGY

In this study, it is proposed to carry out an investigation of the literature related to the mining of educational data in higher education, specifically in works with a predictive approach, and to identify the most used preprocessing procedures and which performance improvements. At the same time, it seeks to examine which performance metrics are being used to assess the results achieved. To this end, a systematic literature review (SLR) methodology will be used, using the guidelines proposed by Kitchenham et al. (2009), which determine three main steps: planning, conducting, and reporting.

Planning

Planning is the first step of SLR and is subdivided into five other subcategories:

Identification of the research objective and questions: The aim of this study is to conduct an investigation of the literature on the use of educational data mining in higher education using the guidelines for systematic review by Kitchenham et al. (2009). To achieve all adjacent purposes, the questions were defined as follows:

Question 1: What preprocessing procedures are described?
Question 2: What attribute selection techniques are described?
Question 3: Which evaluation metrics of the models generated in educational data mining are described?

Identification of keywords: The construction of the search string was elaborated with the concern of properly delimiting the search results without increasing the amount returned in a way that makes the review impossible but also without restricting it too much. In this way, we sought to restrict specific educational data mining jobs used in higher education in the search string. More specific delimitations were carried out in conducting the research through the inclusion and exclusion criteria. The search string used was “Educational Data Mining” AND (“higher education “OR “under graduation courses”).

Identify the sources: The databases used were chosen according to their range of indexing and recognized reliability, delimiting the search to the Scopus, Science Direct and Web of Science databases.
Identify the inclusion/exclusion criteria: Table 1 presents the exclusion and inclusion criteria. It is reiterated that failure to meet the inclusion criteria also excludes the study.

Table 1: Inclusion and Exclusion Criteria

INCLUSION	**EXCLUSION**
Only articles published in journals	Duplicate articles
Published from 2016 to 2021	Articles that present systematic reviews and systematic mappings
Only in English language	Articles that do not have open access
Peer-reviewed journals	Not considering the context of higher education
Explain about all the information inherent to the research questions	Articles not available in PDF format
Studies that implement predictive methods	Articles that do not describe the information inherent to the research questions

Identify the data extraction strategy: in the first step, all articles returned by the search were collected. Second, a transversal reading of the articles was carried out, comprising the title, abstract, methodology and results, excluding those identified within the exclusion criteria and identifying duplicate articles. The remaining articles were read in full, examining eligibility criteria that included meeting the inclusion criteria and a complete description of the study's methodological processes.

CONDUCTING THE REVIEW
It consists of the second stage of the review, subdivided into 5 other stages:
Identify the Search: Currently, searches were carried out in the selected libraries with the search string created. As a result, 95 articles from the Web of Science database, 65 from the Scopus database and 61 from the Science Direct database were obtained, totaling 221 articles.

Selection of studies: As a protocol for examining the collected articles, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol was used (PAGE et al., 2021). With this, it was possible to structure the information flow of the review process, mapping the number of articles identified, included, excluded and the reasons for exclusions. Figure 1 depicts the details of the process of selecting works based on the PRISMA protocol.

Assess the quality of the study: although provided separately in the guidelines of Kitchenham et al. (2009), in this study, the quality was based on the requirements stipulated in the step “Identify the data extraction strategy” and apply it in the PRISMA protocol.

Data extraction: In this step, the data necessary to answer the research questions are extracted, as well as important data to characterize the sample of selected articles. These data are presented and discussed in the results section.
Synthesizing the Data: In this step, the congruence between the analyzed studies and the divergences are highlighted. It mainly seeks to explain the contributions that the studies add to the preprocessing procedures in educational data mining and to the performance metrics used.

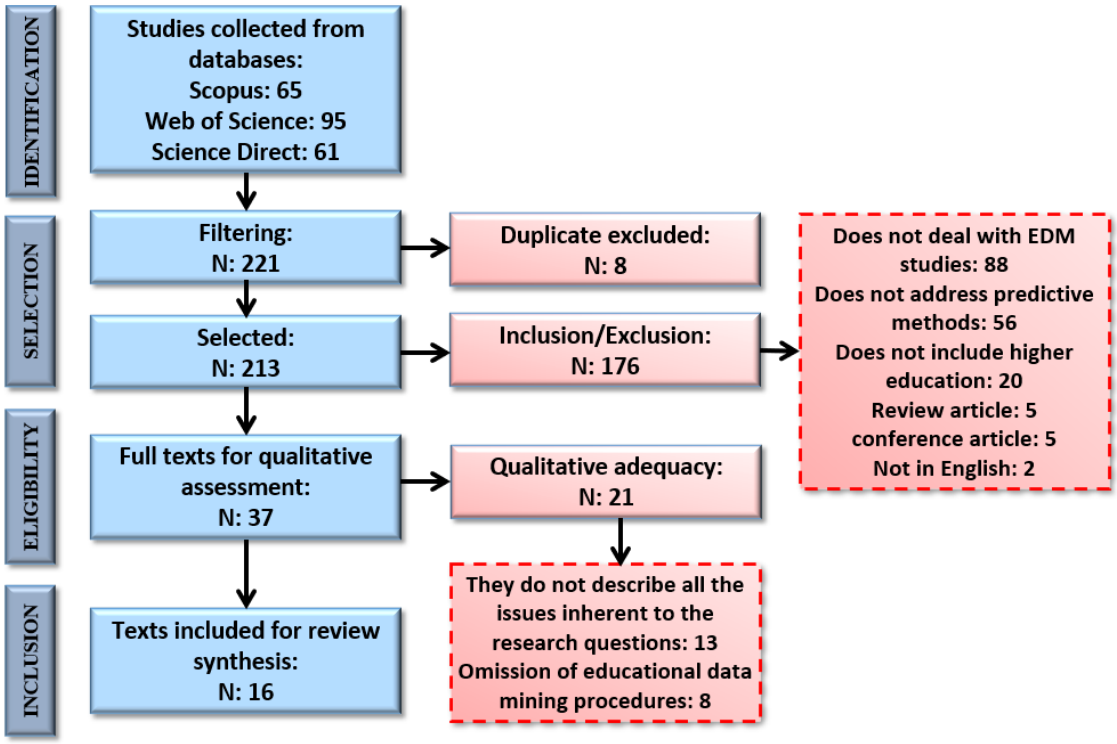

Figure 1: Systematic review protocol based on PRISMA.

RESULTS AND DISCUSSIONS

At this stage, the results are presented both relating to the stipulated research questions and other relevant aspects found in the review. Although specific objectives to be achieved are defined, which lead to an emphasis on preprocessing and metrics, during the full examination of the selected studies, several pieces of information were categorized as a way of characterizing the sample. Regarding the objectives of the studies, there was a predominance of studies on predicting performance in the course, with 10 works on this topic (AKMEŞE, 2021; ANOOPKUMAR, 2018; HELAL, 2018; MIGUÉIS, 2018; ASIF, 2017; SADIQ, 2019; MAGBAG, 2020; COSTA, 2017; ALTURKI, 2021; SULTANA, 2019), one of the works addressing performance in a specific programming discipline (SUNDAY, 2020). Another three works address the prediction of dropout and graduation in courses (AGRUSTI, 2020; MENGASH, 2020; NIETO, 2019; URBINA-NÁJERA, 2020). Two works with purposes that are not so frequent stand out, such as Mengash (2020), who proposes the use of educational data mining to support the university admission process based on the profile and performance obtained in high school, and Pabreja (2017), who used data from every degree to predict the employability potential of its graduates. Regarding the sources for obtaining data for mining, the hegemony of academic systems remains as providers of data for analysis, being indicated by 12 works in the sample of this study, while another three used data from virtual learning environments, and 1 used data from the national education system. However, although several works use the academic system to collect data, it does not mean that the predictor attributes used are the same, as this depends on the researcher's decision or methodology and on the type of information stored in the academic systems of each institution.

When exploring the attributes used in the reviewed studies, classifications were used in categories based on the segmentation performed by Shahiri et al. (2015). Table 2 presents the categories and the description of the attributes that compose them, correlated with the respective works that use attributes of the category.

Table 2: Predictive Attributes

Categories	Category Attributes	Works	The amount
External evaluation	High School Score, Credits from Other Institutions, Admission Score, Subject Specific Entrance Exam Score, GPA High School	(AGRUSTI, 2020), (AKMEŞE, 2021), (MENGASH, 2020), (MIGUÉIS, 2018), (ASIF, 2017), (MAGBAG, 2020), (ALTURKI, 2021)	7

Internal Assessment	Academic Phase, Exam Name, Credits Completed, Maximum Score of an Exam, Academic Phase of the Exam, Result (Pass, Fail), Attendance (Attendance), Average by Academic Phase, Frequency by Academic Phase, Academic Year of Enrollment, Grades in specific subjects, Subjects that failed, Subjects that passed, Maximum Grade, Minimum Grade, Median Grade, Number of Enrollments per term, Grades in Specific Exams, Laboratory Grades, Grade Point Average (GPA), GPA per term, Percentage of Grade Use per academic period.	(PABREJA, 2017), (SUNDAY, 2020), (ALTURKI, 2021), (MAGBAG, 2020), (SADIQ, 2019), (NIETO, 2019), (ASIF, 2017), (MIGUÉIS, 2018), (HELAL, 2018)	9
Demographic data	Age, Gender, Family Income, High School Origin, Parents' Education, Employment Status, Parents' Occupation, Disability, Dependence on Parents or Guardians, Marital Status, Number of Children, Education and Occupation of an Elder Brother	(AGRUSTI, 2020), (AKMEŞE, 2021), (ANOOPKUMAR, 2018), (HELAL, 2018), (MIGUÉIS, 2018), (COSTA, 2017), (URBINA-NÁJERA, 2020), (PABREJA, 2017)	8
Psychometric	Parent Response Survey, Parent School Satisfaction	(SULTANA, 2019)	1
Interaction Data	Book View, Activity View, Course Page View, Forum Post, Start Forum Discussion, Review Forum View, Lesson View, Quiz View, Quiz Take, Quiz Review, Wiki View, Edit wiki, view resource files	(HELAL, 2018), (SADIQ, 2019), (COSTA, 2017)	3

There is a great diversification of attributes, which requires expertise for the analyst on the topic to be addressed to select the attributes, as well as the use of attribute selection techniques that support them. The predominance of the use of all student performance records to make future predictions, considering previous scores, during and after admission to higher education stands out. It may be intuitive that it is expected that a certain performance that the student has at some point will be repeated in future situations. However, it must be considered that, regardless of the level of performance presented by an individual in high school, it was enough for him to pass the subjects and complete this training stage. The same perspective is repeated for the entrance exam because he was approved and joined the institution. What is now discussed is how these performances, which were satisfactory, can now indicate future academic performance. In this context, data mining algorithms may be able to capture the nuances of these relationships. Furthermore, it is possible to combine attribute categories. As shown in Table 2, there are works that describe using attributes from more than one of the categories.

Research Question: What preprocessing procedures are implemented?

Traditional data mining algorithms are rarely applied directly to raw source data, requiring specific preprocessing for the mining methods to be applied (DUTT, 2017). Schmidhuber (2015) argues that preprocessing and data transformations are a byproduct of the data mining strategy, contributing to the representation of data to explore the advantages of the algorithms and minimize the disadvantages. In Han et al. (2011), preprocessing is defined as the process of manipulating, enriching, reducing, or transforming the original data to make them more easily accessible. Urbina-Nájera (2020) complements this understanding by describing that transformation may involve combining data from different sources to provide a unified view. These concepts of preprocessing and data transformation are part of the well-known knowledge discovery process in a database (KDD), comprising as distinct steps the selection of attributes, preprocessing (such as data cleaning) and transformation, all of which are predecessors to the application phase of mining algorithms. In this study, the approach of Sokkhey (2020) is used,

considering as preprocessing all the treatments performed on the data before the application of the algorithms, except for the use of specific techniques for selecting attributes that, by composing a research question, will be addressed in a specific section.

All papers reviewed, at some level, described preprocessing. First, regarding data cleaning, although it is presumable that it is a step present in any EDM enterprise, only half of the reviewed works described performing this activity (AGRUSTI, 2020; AKMEŞE, 2021; ANOOPKUMAR, 2018; MENGASH, 2020; NIETO, 2020; 2019; MAGBAG, 2020; COSTA, 2017; SULTANA, 2019), basically reporting noise elimination, removal of instances with missing data, removal of students who did not fit into any of the predictive classes, removal of duplicates, and fits in this step, the removal of visibly irrelevant attributes such as address, ID, email, number of documents, among others of this nature.

In addition to mandatory data cleaning, the most declared preprocessing procedure is data transformation, also described in 8 works. Regarding performance-related attributes, the need to categorize scores into concepts such as A, B, C, Bad, Good, and Excellent, among others, was demonstrated (ANOOPKUMAR, 2018; MENGASH, 2020; HELAL, 2018; MIGUÉIS, 2018; SUNDAY, 2020; PABREJA, 2017). Age is also an attribute that is usually transformed, especially in age groups. Among the works reviewed, three performed discretization of student ages (AGRUSTI, 2020; HELAL, 2018; PABREJA, 2017). Additionally, in terms of transforming continuous data into discrete data, binarization is also a resource for data transformations and was implemented by Agrusti (2020) in all its attributes. However, the author does not specify exactly the advantages he intended to achieve, only citing "better adequacy of the algorithms" and does not discuss in the results whether advantages were obtained, since there was no comparative dataset (without making the values binary), as it was not an object of study, but only an aspect methodological.

Regarding the transformation, the procedures found and presented here do not have innovative characteristics that aim to advance the mining process. Standard procedures were shown, almost mandatory, to be performed in data preparation for EDM. Two aspects can be highlighted about the methodologies found: first, the approach of transforming scores into concepts. Seven works described carrying out this transformation under the approach of "categorization", while one of them approached this activity as "discretization", which leads us to the second aspect to be highlighted. Although discretization is a common procedure to be performed in data preparation, it is not a random action and, depending on how it is performed, it can negatively influence the prediction. In short, discretization is the division of something into smaller, less complex parts. For this, statistical principles are used so that the new parts faithfully represent the original data.

That said, it appears that the works that described the transformation of scores as categorization did not report using any discretization method. Thus, it is assumed that this was carried out based on the experience of the researchers or on some standard of concepts of the institution. Only Miguéis (2018) used the binning algorithm (equal width binning) to establish five levels of academic performance (A, B, C, D and E) and thus described this transformation as discretization. The importance of formalizing this process with a methodology is highlighted, as generating intervals without observing the amplitude of the sample that one has can generate irregular intervals. It is possible to accumulate, for example, empty age ranges rather than others with more than 60% of individuals in your sample. Table 2 shows that half of the studies used demographic data such as age and salary income range (individual or family). These types of data can have an entirely different amplitude in each context. Therefore, they would need a statistical evaluation for their discretization; however, none of the studies reported using any formal methodology for these attributes.

Still in critical observance of the nonuse of statistical formalization in data processing in preprocessing. It should be noted that several studies report using external scores (high school grades and entrance exam grades) and internal scores (semester grades, grades in subjects or exams and activities) in the same dataset for data mining. Only Magbag (2020) reported the concern and need to perform data normalization. This aspect is relevant, as the scoring scales may differ from one data source to another.

A very common situation in data mining is unbalanced databases, that is, the number of instances for each predictive class is too discrepant. This is a very plausible scenario to occur for several reasons in educational databases. In exact science courses, for example, the number of students who drop out or fail may be much higher than in other possible situations for students. The same scenario can be repeated in courses or disciplines, traditionally with few failures or dropouts. In addition to harming the construction of the prediction model, this situation can lead to a misinterpretation of the results and the effectiveness of the algorithm. Take as an example a subject in which the approval rate is 90% (in the training group, 90% are approved), if this algorithm correctly classifies the prediction of approved students at 90% (the tendency is to print a predominance of correct answers

in the majority class) and miss 100% in the prediction of students at risk of failing, the accuracy of this algorithm will still be 81%, which is a good success rate, but, in practice, this prediction model is useless, since that the objective is to highlight students at risk of failing so that it is possible to carry out some form of intervention, but in this case, the minority class will always be misclassified.

Only three works reported the observation of this problem and implemented some load balancing technique, namely, MAGBAG (2020), COSTA (2017), and ASIF (2017). Of these, only Magbag (2020) and Costa (2017) specified which technique they used; both works used the Synthetic Minority Oversampling Technique (SMOTE) algorithm that works by creating synthetic minority class data based on existing neighbors. Assif (2017) does not describe which load balancing technique was used, but compares the results with the unbalanced base, reporting that there were no improvements in classification rates, but does not discuss the confusion matrix, so it is not known if there were any changes in the classification of minority classes.

Are attribute selection techniques being used? What techniques?
As shown in Table 2, there is a significant diversification of attributes present in the consulted databases. However, not all of them may be relevant to the characterization of the desired prediction state. In fact, using all available attributes can be harmful. As described by Urbina-Nájera (2020), many attributes represent a large dimensional space, and it is necessary to perform a dimensionality reduction, selecting only a few attributes in a way that retains as much information as possible to describe the training instances.

In some cases, only the researcher's expertise around analysis, experience in data mining or knowledge of the algorithmic techniques to be applied are used in the selection of attributes. However, there are techniques whose objective is to select the most relevant factors to be used as input variables for forecast models (ALBÁN, 2019). Among the works examined, seven reported using some procedure for selecting attributes (AKMEŞE, 2021; ANOOPKUMAR, 2018; MENGASH, 2020; MAGBAG, 2020; COSTA, 2017; ALTURKI, 2021; URBINA-NÁJERA, 2020). With emphasis on the use of the GainRatioAttributeEval method, which evaluates each attribute by measuring its proportion of influence related to the class, and the InfoGainAttributeEval method, which estimates the attributes by measuring the information gain provided by each attribute related to its class, implemented by 4 works (MAGBAG, 2020; COSTA, 2017; ALTURKI, 2021; URBINA-NÁJERA, 2020). The other methods used were the chi-squared method (AKMEŞE, 2021) and the correlation coefficient method (MENGASH, 2020). Anoopkumar (2018) does not report which technique was used but states that an attribute selection process was carried out.

Among the results obtained with the selection of attributes, Alturki (2021) shows that the student's cumulative average (GPA) for each semester, the number of courses failed in the 1st and 2nd year, his grade in the 'Fundamentals of the Database' and 'Basic programming courses 1' are the most influential attributes for predicting academic performance, while English skills and high school cumulative average have not been shown to influence the prediction. Urbina-Nájera (2020) reduced its attributes from 56 to 27 after using the selection method, and Akmeşe (2021) reduced it from 12 to 6 attributes. In both cases, the attributes of gender and age were waived. Alturki (2020) reported in his investigation that gender and age did not show significant impacts in his reviewed studies, but they are among the attributes most used by most studies. Table 2 corroborates this finding. Alban (2019) attributes the recurrence of the use of these attributes because they are internal data that are simple to define and measure; that is, they are easily found in academic systems, teaching platforms or sources of demographic data. In this way, gender and age are trivial information that arise in any collection performed.

What evaluation metrics are being used to evaluate results in educational data mining studies?
It is part of the analysis of results in data mining to evaluate the classification model, which allows the researcher to determine the level of confidence in the prediction model, which supports the interpretation of the results and, if necessary, guides the adjustments and repetition of the preprocessing. Souza (2021) argues that to verify the results of a classification model, it is necessary to define the evaluation methods and the interpretation metrics because together they can determine whether a model is effective or not. A diverse set of metrics are available to researchers, but in general, the confusion matrix is the dominant measure, and its components are used in the calculation of many metrics (SHUQFA, 2019).

Usually, cross validation is predominant as an evaluation method. This method performs the random division of the dataset for the algorithm evaluation in equally distributed data subsets. After the division, the tests are performed by sending one of the subsets for the algorithm to classify (test set), and the others act as a training set. The process is repeated until all subsets have been used as a test set and a training set. In this survey, this was the method used in 13 studies. Only Agrusti (2020), Sunday (2020) and Anoopkumar (2018) did not specify which evaluation method was used.

For the metrics used, a variety of 8 metrics were found. Table 3 presents the studies examined and the metrics used to evaluate the results. The greater use of the metrics of accuracy, precision, and F measure was reported in 10 works, followed by recall, which was reported in 9 works.

Table 3: Evaluation Metrics

Studies	Metrics							
	Accuracy	**Precision**	**F Measure**	**Recall**	**Specificity**	**Kappa**	**AUC**	**ROC**
(AGRUSTI, 2020)		X	X	X				
(AKMEŞE, 2021)	X							
(ALTURKI, 2021)		X	X					
(ANOOPKUMAR, 2018)	X	X	X	X	X			
(ASIF, 2017)	X					X		
(COSTA, 2017)			X					
(PABREJA, 2017)		X						
(HELAL, 2018)		X	X	X		X	X	
(MAGBAG, 2020)	X	X	X	X			X	
(MENGASH, 2020)	X	X	X	X				
(MIGUÉIS, 2018)	X	X		X				
(NIETO, 2019)							X	
(SADIQ, 2019)	X		X	X				
(SULTANA, 2019)	X				X	X		X
(SUNDAY, 2020)	X	X	X	X		X		
(URBINA-NÁJERA, 2020)	X	X	X	X				

In addition to the diversity of metrics, the survey shows that, except for Nieto (2019) and Pabreja (2017), all other works employ more than one metric. However, it is essential to understand to what extent the characteristics of each of these metrics are being considered for interpreting the results and evaluating the forecast model. Table 4 displays the description of each of the metrics recorded.

Table 4: Description of Evaluation Metrics[1]

Metrics	**Description**
Accuracy	Proportion of samples correctly estimated in relation to the number of all samples. That is, the test is the rate of total correct diagnoses.
Precision	Ratio between true positives and all values classified as positive.

[1] For a better understanding of the descriptions in this table, consider: True Positive Rate (TP): the number of instances that are correctly predicted as positive; False Positive Rate (FP): The number of occurrences that are incorrectly predicted to be positive; True Negative Rate (TN): the number of instances that are correctly predicted to be negative; False Negative Rate (FN): The number of instances that are incorrectly predicted to be negative.

Recall	Corresponds to the success rate of positive examples correctly classified among all positive examples in the base.
F Measure	It represents the harmonic mean between recall and precision.
Specificity	It is the true negative rate, that is, it is the number of instances of the negative class that were predicted to be negative.
Kappa	It compares the overall accuracy of the classification model with the expected accuracy for the same if the classification is performed at random.
ROC	It consists of a two-dimensional graph, where the y-axis refers to Precision or Recall and the x-axis to specificity.
AUC	Synthesizes the ROC curve into a single value, aggregating all ROC thresholds, calculating the "area under the curve".

It is evident that although some metrics complement each other, others deal with specific results of the classification. The finding made in this investigation about the lack of description or omission of the criteria to use each metric, and how they were used to interpret the results or evaluate the forecast model, is worrisome. Although they use different metrics, when presenting the results of the best forecasting techniques, the result obtained by accuracy and, lacking it, precision is used. Only Helal (2018) considers the best forecast model based on the F-measure metric.

As described in Table 4, the accuracy represents the overall performance of the forecast model. Although it is the most commonly used metric in data mining studies, it must be used and interpreted carefully. Otherwise, it can lead to a mistake in the evaluation of the model, especially in cases of unbalanced classes. A high accuracy rate may be based on the hits of the majority classes, neglecting minority classes. Regarding educational data, the bases are likely to be unbalanced in different ways, depending on the institution, course, discipline, etc. Thus, the need to explore additional metrics is evident.

The Kappa coefficient is a metric whose approach seeks to mitigate this situation, printing a more reliable performance metric, regardless of the base imbalance. With these characteristics, it is widely used to determine the confidence level of forecast models. However, the objectives and context of the problem to be addressed with educational data mining may require more specific metrics, which can support the analyst and obtain more satisfactory results. The Precision, Recall, and Specificity metrics are highlighted here, as they are metrics through which it is possible to separately assess the performance of the specific forecast for each class.

The objective of all 16 works reviewed using educational data mining was to build a predictive model of the results of students in courses or subjects that can be interpreted as positive or negative situations. Positive, for example, in cases of graduation, approval or performance concepts Good, Excellent, A or B. Negative, in cases of evasion, retention, failure, or performance concepts Bad, Terrible, C, D or E. Even the objectives of Pabreja (2017) about the employability or not of graduates fit into predictions of positive cases (in case of the possibility of getting a job) and negative cases (in case of not getting a job). With this, it is evident that the benefits of predicting these results lie in the possibility of carrying out an early intervention in students at risk of negative results to change their trajectories, transforming possible negative trends into positive results.

In this way, it is preferable that the prediction models have an excellent performance in identifying students at risk (with prediction of dropout, retention, failure, low performance, or unemployment) so that they can be contemplated by some intervention action. Wrongly predicting a student who would be approved as a student at risk of failing will cause him to receive some pedagogical intervention to improve his performance, which will not cause him any harm; at most, this situation can lead to an unnecessary allocation of resources. However, a student who is truly at risk of failing and is classified as a student expected to perform well and pass will not receive any intervention and may continue his tendency to fail.

Results obtained by the classification techniques used

Although it is not part of the research questions of this work, the results achieved by the data mining techniques used were extracted during the examination in full of the studies. A total of 21 different algorithms were identified; however, the predominance of decision tree techniques (used in 15 works), naïve Bayes (used in 14 works), forests and neural networks (both used in 7 works) stands out.

In addition to the predominance in its implementation, these techniques presented the best performances. The decision tree technique presented the best performance in 6 works, followed by Forests, which stood out in 5 works. Naïve Bayes had the best performance in 3 works, and finally, Neural Networks obtained the best results in 2 works. Only in Helal (2018) was the best result presented by a technique outside this group, using rule-based mining, specifically the JRIP algorithm.

CONCLUSIONS

The present study carried out a systematic review of the literature on educational data mining used in predictive approaches, covering only higher education. Within the delimited specifications, the superiority of studies aimed at building a model for predicting academic performance was evidenced. However, other forecasting possibilities were implemented, demonstrating that there is flexibility for the use of EDM if the necessary information is available. In this regard, the feasibility of performing educational data mining is confirmed since institutions already produce some daily information that can be processed and mined. Most studies reported using data from academic systems and teaching platforms.

This study had an initial premise that EDM studies tend to use several algorithms in the same research, comparing the results at the end. This was confirmed, and 21 different algorithms were identified, distributed in studies that implemented 3 to 10 algorithms at the same time. This justifies the objectives of this work and seeks to bring a new contribution, examining the preprocessing procedures performed, understanding that the preprocessing and selection of attributes have as much influence on obtaining good results in EDM as the algorithms used.

After the review is carried out, first, the analyses demonstrate that more requirements must be added to the work of the researcher who employs EDM, and the need to use several formal and statistical methods in the selection of attributes and preprocessing of the data is recommended. However, it should be borne in mind that, in the same way that the use of different algorithms can occur due to the ease offered by the tools that researchers have at their disposal, this could reflect on preprocessing activities. All works reported using support tools, namely, Weka, RapidMinder, Pentaho and libraries of Python and R languages. These same tools have resources available for attribute selection, balancing, normalization, discretization, among other features. All reported preprocessing methods can be performed by the same tools that were used to apply the algorithms.

The examination, carried out to answer the research question relating to metrics, clarified that this topic needs to be discussed in more depth. Several studies reported using 1 to 5 metrics, but it was not appreciated to what extent each of them would be suitable for each scenario. This finding shows that, in the same way that the tools bring many algorithms, they generate many metrics as results, and all of them are brought as “search results”. However, the purposes for the use of each metric specifically are absent in the studies.

As future aspirations, the need for studies to compare or validate preprocessing techniques and understand the results that metrics can represent in educational data mining is considered. It is considered that the results of studies in this sense are likely to be generalized or extended, as there is an equivalence of data sources as well as the attributes used by studies on this topic.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Abu Saa, A., Al-Emran, M. & Shaalan, K. (2019). “Factors Affecting Students’ Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques. Tech Know Learn 24, 567–598. https://doi.org/10.1007/s10758-019-09408-7

Ackcapinar, G., G, M. N. Hasine, R. Majumdar, B. Flanagan, and H. Ogata. (2019). “Developing early system for spotting at-risk students by using eBook interaction logs,” Smart Learning Engineering, vol. 6, Issue 4, pp. 1-15, March.

Agrusti, F., Mezzini, M., and Bonavolontà, G. (2020). “Deep learning approach for predicting university dropout: a case study at Roma Tre University”. Journal of E-Learning and Knowledge Society, 16(1), 44-54. https://doi.org/10.20368/1971-8829/1135192

Akmeşe, Ö. F., Kör, H., and Erbay, H. (2021). “Use Of Machine Learning Techniques For The Forecast Of Student Achievement In Higher Education”. Information Technologies and Learning Tools, 82(2), 297–311. https://doi.org/10.33407/itlt.v82i2.4178

Albán, M. and Mauricio, D. (2019). “Predicting University Dropout through Data Mining: A Systematic Literature. Indian Journal of Science and Technology”. Journal of Science and Technology. Volume:

12, Issue: 4.
Alturki, S., Hulpuș, I. and Stuckenschmidt, H. (2020). “Predicting Academic Outcomes: A Survey from 2007 Till 2018.” Tech Know Learn. https://doi.org/10.1007/s10758-020-09476-0
Alturki, S., and Alturki, N. (2021). “Using educational data mining to predict students’ academic performance for applying early interventions”. Journal of Information Technology Education: Innovations in Practice, 20, 121- 137. https://doi.org/10.28945/4835
Anoopkumar, M. and Zubair Rahman, A. M. J. Md. (2018). “Bound Model of Clustering and Classification (BMCC) for Proficient Performance Prediction of Didactical Outcomes of Students” International Journal of Advanced Computer Science and Applications (IJACSA), 9(11). http://dx.doi.org/10.14569/IJACSA.2018.091133
Asif, Raheela, Agathe Merceron, Syed Abbas Ali, Najmi Ghani Haider. (2017). “Analyzing undergraduate students' performance using educational data mining”. Computers & Education, Volume 113, Pages 177-194, ISSN 0360-1315, https://doi.org/10.1016/j.compedu.2017.05.007.
Costa, Evandro B., Fonseca, Baldoino., Santana, Marcelo Almeida., Araújo, Fabrísia Ferreira and Rego, Joilson. (2017). “Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses”. Computers in Human Behavior. Volume 73, Pages 247-256, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2017.01.047.
Dutt A., Ismail M. A. and T. Herawan. (2017). "A Systematic Review on Educational Data Mining," in IEEE Access, vol. 5, pp. 15991-16005, doi: 10.1109/ACCESS.2017.2654247.
Helal, S., Jiuyong Li, L. L., Ebrahimie, E., Dawson, S., Duncan J. Murray, Qi Long. (2018). “Predicting academic performance by considering student heterogeneity. Knowledge-Based Systems”, Volume 161, Pages 134-146, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2018.07.042.
Magbag, A., and Raga, R.C. (2020). “Prediction Of College Academic Performance of Senior High School Graduates Using Classification Techniques”. International journal of scientific & technology research volume 9, issue 04, April.
Mengash, H. A., (2020). "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," in IEEE Access, vol. 8, pp. 55462-55470, doi: 10.1109/ACCESS.2020.2981905.
Miguéis, V. L., Freitas, Ana., Garcia, Paulo J.V. Silva, André. (2018). “Early segmentation of students according to their academic performance: A predictive modeling approach”. Decision Support Systems, Volume 115, Pages 36-51, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2018.09.001.
Nieto, Y., Gacía-Díaz,V., Montenegro, C., González, C. C. and González Crespo, R., (2019). "Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions," in IEEE Access, vol. 7, pp. 75007-75017, doi: 10.1109/ACCESS.2019.2919343.
OECD. (2019). “Higher Education in Mexico: Labor Market Relevance and OutcEomes”, Higher. París: OECD Publishing. https://doi.org/10.1787/9789264309432-em
Pabreja, K., (2017). “Comparison of Different Classification Techniques for Educational Data,” International Journal of Information Systems in the Service Sector (IJISSS), IGI Global, vol. 9(1), pages 54-67, January. http://doi.org/10.4018/IJISSS.2017010104
Page, M.J., Mackenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al., (2021). “The prisma 2020 statement: an updated guideline for reporting systematic reviews”. Bmj 372.
Peña-Ayala, A. (2014). “Educational data mining: A survey and a data mining-based analysis of recent works”. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2013.08.042.
Sadiq, M. H. and Ahmed, N. S. (2019). “Classifying and Predicting Students’ Performance using Improved Decision Tree C4.5 in Higher Education Institutes”. Journal of Computer Science, 15(9), 1291-1306. https://doi.org/10.3844/jcssp.2019.1291.1306
Schmidhuber, J. (2015). “Deep learning in neural networks: An overview. Neural Networks”, Vol 61, pp 85-117, Jan 2015.
Shahiri, A. M., Husain, W., and Rashid, N. A. (2015). “A review on predicting student’s performance using data mining techniques”. Procedia Computer Science. https://doi.org/10.1016/j.procs.2015.12.157.
Shuqfa, Z. and Harous, S. (2019). "Data Mining Techniques Used in Predicting Student Retention in Higher Education: A Survey," International Conference on Electrical and Computing Technologies and Applications (ICECTA), 2019, pp. 1-4, doi: 10.1109/ICECTA48151.2019.8959789.
Shahiri, A.M., Husain, W. and Rashid, N.A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science, 72, 414-422. International Journal of Information Systems in the Service Sector Volume 9, Issue 1, January-March 2017
Sokkhey, P. and Okazaki, T., (2020). “Developing Web-based Support Systems for Predicting Poor-performing Students using Educational Data Mining Techniques” International Journal of Advanced Computer Science and Applications (IJACSA), 11(7), http://dx.doi.org/10.14569/IJACSA.2020.0110704

Souza, V. F., Santos, T. C. B. (2021). “Processo de Mineração de Dados Educacionais aplicado na Previsão do Desempenho de Alunos: Uma comparação entre as Técnicas de Aprendizagem de Máquina e Aprendizagem Profunda.” Revista Brasileira de Informática na Educação, [S.l.], v. 29, p. 519-546, jun. 2021. ISSN 2317-6121.

Sultana, J. & Rani, M. and Farquad, H. (2019). “Student’s Performance Prediction using Deep Learning and Data Mining methods”. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June.

Sunday, K., Ocheja, P., Hussain, S., Oyelere, S.S., Samson, B.O., and Agbo, J.F. (2020). “Analyzing Student Performance in Programming Education Using Classification Techniques”. iJET, 15, 127-144.

Urbina-Nájera, A.B., Camino-Hampshire, J.C., & Cruz-Barbosa, R. (2020). “University dropout: Patterns to prevent it by applying educational data mining”. RELIEVE, 26(1), art. 4. http://doi.org/10.7203/relieve.26.1.16061