Technology, Knowledge and Learning
https://doi.org/10.1007/s10758-019-09408-7

ORIGINAL RESEARCH

Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques

Amjed Abu Saa[1] · **Mostafa Al-Emran**[2] · **Khaled Shaalan**[1]

© Springer Nature B.V. 2019

Abstract
Predicting the students' performance has become a challenging task due to the increasing amount of data in educational systems. In keeping with this, identifying the factors affecting the students' performance in higher education, especially by using predictive data mining techniques, is still in short supply. This field of research is usually identified as educational data mining. Hence, the main aim of this study is to identify the most commonly studied factors that affect the students' performance, as well as, the most common data mining techniques applied to identify these factors. In this study, 36 research articles out of a total of 420 from 2009 to 2018 were critically reviewed and analyzed by applying a systematic literature review approach. The results showed that the most common factors are grouped under four main categories, namely students' previous grades and class performance, students' e-Learning activity, students' demographics, and students' social information. Additionally, the results also indicated that the most common data mining techniques used to predict and classify students' factors are decision trees, Naïve Bayes classifiers, and artificial neural networks.

Keywords Educational data mining · Students' performance · Data mining techniques · Systematic review

1 Introduction

An increasing interest has arisen during the past decade to identify the most important factors influencing students' performance in higher education, especially by using data mining methods and techniques. This field of research is usually identified as educational data

✉ Mostafa Al-Emran
al.emran@tdtu.edu.vn

Amjed Abu Saa
a.abusaa@ajman.ac.ae

Khaled Shaalan
khaled.shaalan@buid.ac.ae

[1] Faculty of Engineering and IT, The British University in Dubai, Dubai, UAE

[2] Applied Computational Civil and Structural Engineering Research Group, Faculty of Civil Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Published online: 25 April 2019

mining (EDM) (Bakhshinategh et al. 2018). The motivation behind this interest is attributed to the applicability of such research in helping to identify low performing students early enough to overcome their difficulties in learning and improve their learning outcomes, which in turn serves the institutional goals of providing high-quality education ecosystems. In addition, EDM is fast becoming an important field of research due to its ability to extract new knowledge from a huge amount of students' data (Wook et al. 2017). This paper is equally interested in this topic, and our objective is to explore and review papers from the past decade that are in the context of educational data mining and identifies the main factors influencing students' performance in higher education. EDM is defined by the Educational Data Mining community website (www.educationaldatamining.org) as "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". In general, EDM is a set of methods that apply data mining techniques such clustering, classification, prediction among the others to the data retrieved by many educational systems (Berland et al. 2014).

Researchers in this field tend to study different types of students' factors and attributes that affect students' performance and learning outcomes (Abu Saa et al. 2019). Shahiri et al. (2015) conducted a systematic literature review on predicting students' performance using data mining techniques. The review tackled many subjects, one of which was to identify the important attributes used in predicting the students' performance. Results showed that the cumulative grade point average and internal assessments are the most frequent attributes used for predicting the students' performance. Furthermore, other important attributes were also identified, including students' demographic and external assessments, extra-curricular activities, high school background, and social interaction network. Additionally, the results showed that the Decision tree and Neural Networks were the most frequently used data mining techniques for predicting students' performance. Besides, Peña-Ayala (2014) conducted a survey and meta-analysis of research studies related to EDM. The results indicated that 60% of EDM research articles have used predictive data mining approaches as opposed to 40% which have used the descriptive approaches. Furthermore, the results also indicated that classification and clustering were the most typical techniques used by EDM research. Additionally, Bayes theorem, decision trees, instance-based learning (IBL), and hidden Markov model (HMM) were found to be the most popular methods used by EDM research. Furthermore, Romero and Ventura (2007) carried out a review study aiming to analyze the application of data mining for different educational systems: traditional system, web-based courses, content management systems, and intelligent web-based systems. The results suggested investigating the applicability of using data mining techniques for e-learning systems.

Predicting the students' performance has become a challenging task due to the increasing amount of data in educational systems (Shahiri et al. 2015). It is also argued that the existing prediction methods are still insufficient to determine the appropriate techniques for predicting the students' performance in higher educational institutions. Moreover, the identification of the factors affecting the students' performance is still neglecting and requires further research. Thus, there is a clear need to identify the most important factors that were found significant and truly affect the students' performance from the increasing amount of EDM literature. The aim of this paper is to investigate the literature related to EDM and to identify the most important and most studied factors influencing the students' performance in higher education, as well as, to produce a generalized set of factors and attributes that are believed to affect the students' performance and learning outcomes in the higher education sector. In order to perform this study, the existing literature has been reviewed using

a systematic literature review (SLR) method. SLR is one of the most common approaches used for literature review (Al-Emran et al. 2018), and it serves our objective which supposed to provide a summary of studies related to EDM and identify the factors affecting students' performance in higher education.

The rest of the paper is organized as follows. The methodology of the systematic literature review is explained in the next section. Section 3 demonstrates the results achieved from this study. Discussing the results is outlined in Sect. 4, whereas the concluding remarks are demonstrated in Sect. 5.

2 Systematic Review Method

In this paper, we employed a standard SLR methodology, which follows the guidelines proposed by Kitchenham et al. (2009). SLR has many advantages over simple and unstructured literature review methods, as it is more likely to be considered reliable and unbiased (Al-Qaysi et al. 2018). Additionally, information gathered from SLRs is highly reliable as it was derived from various sources (Kitchenham and Charters 2007). There are three main phases of SLR, namely planning, conducting, and reporting (Kitchenham et al. 2009). The first two phases are discussed in the following two sub-sections, whereas the third phase is discussed in Sect. 3.

2.1 Planning

Planning represents the first phase of the SLR method which includes five steps as per the following sub-sections.

2.1.1 Identify the Research Goal and Research Questions

Our objective in this paper is to systematically review relevant literature through the SLR process (Kitchenham and Charters 2007), and our research questions are provided as follows:

- What are the factors affecting students' performance in higher education?
- What are the data mining techniques used to analyze and predict the students' performance?

2.1.2 Identify the Keywords

Our search keywords were mostly driven by the research questions stated in the previous subsection. After identifying the search keywords, we had to prepare a search string that should work with the search engines of the libraries to be searched which will be identified in the next section. The search string used in this study is: [("data mining" OR "educational data mining") AND ("factors affecting students' performance" OR "predicting students' performance")].

As it can be seen in the search string, the term "predicting students' performance" was added to the search string even though it did not appear in the research questions. This is because we have noticed in the planning stage that there are a lot of research studies that included this term in their titles and/or abstracts which indicates that the research is

related to EDM, and these research studies predict the students' performance based on other attributes.

2.1.3 Identify the Sources

The following online library databases and search engines were selected to be searched for the current SLR: ScienceDirect, EBSCO, ProQuest, JSTOR, and Taylor & Francis. The authors of this study assumed that these databases are the main sources for collecting articles related to EDM and students' performance.

2.1.4 Identify the Inclusion/Exclusion Criteria

Our inclusion criteria are shown in Table 1. Each study found in the search results must meet these criteria in order to be included in our SLR.

2.1.5 Identify the Data Extraction Strategy

In this study, the data were collected based on the fields described in Table 2. The research articles that did not have one or more fields from the data described in Table 2 were excluded from the study.

2.2 Conducting the Review

Conducting the review represents the second phase of the SLR method which includes five steps as per the following sub-sections.

2.2.1 Identify the Research

In this step, we started searching the online libraries' databases with the aforementioned search string. The initial search results returned by the search engines are illustrated in Table 3.

2.2.2 Select the Studies

In this study, the selection of articles was carried out according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) (Liberati et al. 2009).

Table 1 Inclusion criteria

Inclusion criteria
a. Must meet the research keywords conditions
b. Must be classified as a data mining or machine learning research
c. Must include the studied factors
d. Full-text articles must be available and accessible, and must not be accessible via arXiv
e. Must not be a review paper
f. Must be published in the last decade (i.e., between 2009 and 2018)
g. Must be written in English language

Table 2 Data layout

Item	Item description
Paper ID	An ID number is assigned to each research paper in order to be easily referenced during the review
Source	The database source of the research paper
Paper title	The title of the research paper
Journal	The journal that published the research paper
Author	The author(s) of the research paper
Year	The paper publication year
Country of study	The country in which the study of the research paper was undertaken
Studied factors	The list of factors that were studied in the research paper
Factors category(s)	The categories of the factors in the previous field, such as: students demographics, students social information, e-Learning activities, etc.
Factors found significant	The list of factors that were found significant by the researchers of this study out of the full list of studied factors
Data mining approaches	The data mining techniques used in the research paper, such as: classification, clustering, etc.
Data mining algorithms	The data mining algorithms that were used in the research paper, such as: decision trees, SVM, K-Means clustering, etc.
Data collection techniques	The techniques that were used to collect data in the research paper, such as: surveys, student information systems data, e-Learning system data, etc.
Dataset size	The size of the dataset that was used in the research paper

Table 3 Initial search results

Online library database/search engine	Results
Science direct	48
EBSCO	80
ProQuest	34
JSTOR	201
Taylor and Francis online	57
Total	420

PRISMA exhibits the flow of information throughout the SLR phases, in which it maps the number of articles identified, included, excluded, and the reasons behind exclusions (Liberati et al. 2009). Figure 1 illustrates the PRISMA flowchart. In that, we select the research articles that satisfy the inclusion/exclusion criteria and examine the contents of the selected articles to verify their eligibility for selection. We then applied the automatic and semi-automatic article selection. The process of automatic selection involves examining the articles against the titles, abstracts, and keywords, whereas the semi-automatic selection involves reading the full-text of the remaining articles.

As a result of the automatic selection, we found 218 duplicate articles, in which, 57 did not meet the research keywords and were not related to the subject of this SLR study, 19 were not a data mining research, 23 were not free access nor available, or an arXiv paper, 21 were review papers, and 34 were non-English papers. Furthermore, research dates were set as a search filter in the search engine and were excluded from the initial search results.

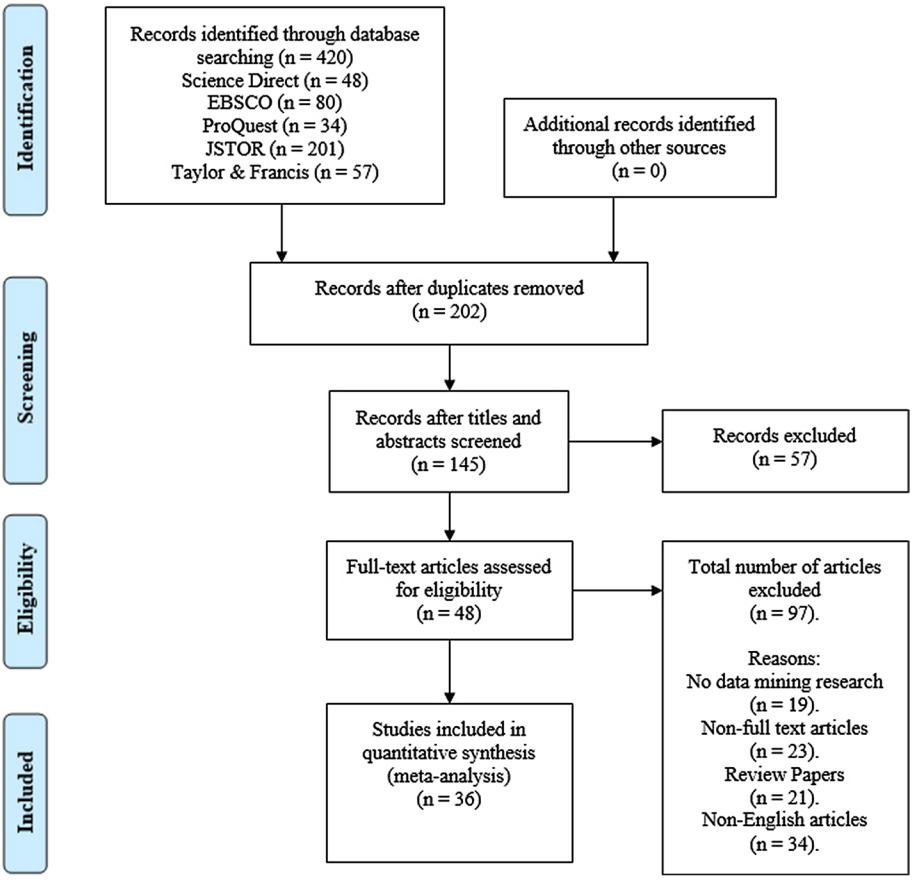

Fig. 1 PRISMA flowchart

Finally, we were left with 48 research papers. By applying the semi-automatic paper selection and reading the full-text articles, it was found that 12 papers did not include the factors that affect the students' performance; hence, they were also excluded from the list. Therefore, the final dataset size of the SLR is 36 research papers. Table 11 (See "Appendix") lists all the 36 research papers selected for this SLR.

2.2.3 Assess the Study Quality

In order to assess the quality of the selected papers in the previous sub-section, we have to answer the questions in Table 4 for each paper in the dataset. It is worth mentioning that the analysis of the collected studies was carried out by the first and second authors of this study by analyzing each article manually. The intercoder agreement rate for coding was 97.5%. The analysis differences between the two authors were resolved through discussion and further review of the disputed studies.

Correspondingly, the answers to the questions in Table 4 accept three scores: Yes (1), Partially (0.5), and No (0). Summing up the scores of all questions for each study will result in a cumulative quality score for each paper out of 9. The result is then converted

Table 4 Quality assessment questions (Kitchenham and Charters 2007)

#	Question
Q1	Are the study aims clearly stated?
Q2	Is the research described adequately?
Q3	Does the study explore diversity of perspectives and contexts?
Q4	Do the objectives lead to conclusions clearly?
Q5	Are the findings important?
Q6	Are negative findings presented?
Q7	Do the researchers explain the consequences of any problems?
Q8	Does the study add to your knowledge or understanding?
Q9	Do the results add to the literature?

into a percentage, e.g., 7 out of 9 is 77.78%. Table 12 (See "Appendix") shows the cumulative quality scores for all the papers included in our dataset along with their percentage results, sorted by percentage.

As shown in Table 12, 34 out of 36 papers achieved a score equal to or greater than 4.5 (50%), whereas only two scored a value of 4/9 (44.44%) due to their low quality and poor content. The most top scored papers are RP10 and RP22 with a score of 8/9 (88.9%). Consequently, the two low-scored papers (RP130, RP136) were removed from the SLR process, and the remaining 34 were kept for the subsequent steps.

2.2.4 Extract the Data

In this step, the required data for our SLR study will be extracted from the selected papers according to the data layout in Table 2. We focused in this step on finding the factors that affect the students' performance, and most importantly, those that were found significant by the researchers in their papers, as well as, the data mining techniques and algorithms used by the researchers in their data mining research. These data will help us to get useful insights and results that will empower us to answer our research questions. Table 13 (See "Appendix") summarizes the extracted data for each paper in our dataset.

2.2.5 Synthesize the Data

We extracted 215 distinct significant factors from the 34 research papers that affect the performance of students in their education life. Furthermore, during the data extraction process, we identified the category for each set of factors that were collected from each article. As a result, we found nine-factor categories that the 215 factors belong to. Table 5 shows an extract of the nine categories of factors along with their descriptions, whereas the sources and number of articles for these categories are illustrated in the extended part of the Table 14 in "Appendix".

3 Results

In this section, we report the results of our SLR study, where we will answer our research questions, and elaborate on the interesting results we came up with from the extracted data.

Table 5 Description of factors' categories (extract)

Category	Description
Students e-Learning activity	The activity logs of students in e-Learning systems, such as, the number of logins, the number of assignments done, number of quizzes done, etc.
Students previous grades and class performance	The grades or other performance indicators of students in previous courses, semesters, or years
Students environment	The attributes of a student environment, such as: the type of school, the type of classroom, class period, etc.
Students demographics	The demographic data of a student, such as: gender, age, nationality, ethnicity, etc.
Instructor attributes	Information about the instructor of the student and his/her evaluation results
Course attributes	Information about the course or module the student is taking, such as, length of the course, difficulty, etc.
Students social information	Information related to the student social life, like the number of friends, if s/he smokes or not, etc.
Course evaluations	Data collected from course evaluation surveys, such as, questions related to the clarity of the course, the level of satisfaction, etc.
Students experience information	Information about the experience of students about the course, such as the readiness of the student, and self-efficacy

3.1 Research Question 1: Factors Affecting Students' Performance

3.1.1 Distribution of Research Articles by Factors' Categories

We classified each article under one or more categories as described in Table 5. These include: (1) Students e-Learning activity, (2) Students previous grades and class performance, (3) Students environment, (4) Students demographics, (5) Instructor attributes, (6) Course attributes, (7) Students social information, (8) Course evaluations, and (9) Students experience information. As it can be seen in Fig. 2, the most common and widely used

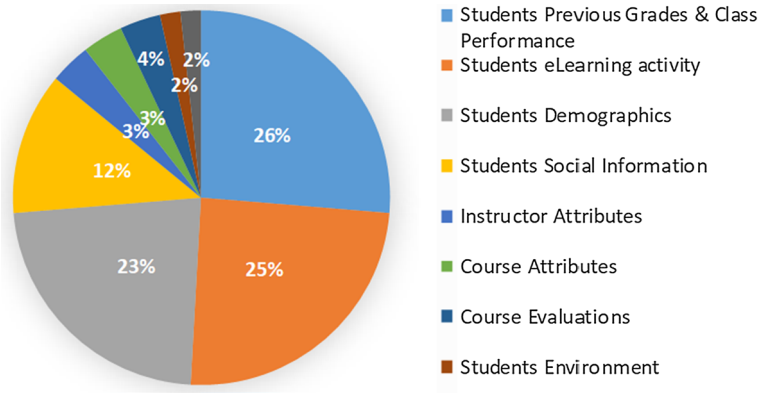

Fig. 2 Distribution of research articles by category

factor categories for predicting students' performance in higher education are students' previous grades and class performance (26%), followed by students' e-Learning activity (25%), students' demographics (23%), and students' social information (12%), respectively. These 4 categories were presented in 86% of the analyzed research studies.

This finding is in agreement with the findings of a prior systematic literature review conducted by Shahiri et al. (2015), which showed that the CGPA and internal assessment marks are the most frequently used attributes in the EDM community for predicting the students' performance. This matches our top factor category which represents the "students' previous grades and class performance". The other five categories which represent a total of 14% altogether were used by a few number of studies and did not frequently appear in other research articles; therefore, they were considered as ad hoc factors.

3.1.2 Distribution of Research Articles by Publication Year

Table 6 indicates that educational data mining research field was most popular in 2016, where more than 17.5% of the research was conducted in this year, and a significant increase of interest was started on 2012.

3.1.3 Distribution of Research Articles by Data Collection Techniques

Another dimension was added to the data collection, which is the techniques of the data collection. In fact, we identified five techniques of data collection, namely: (1) e-Learning system logs, (2) Student information system data, (3) Surveys, (4) Course evaluations surveys, and (5) Network access logs. Table 7 summarizes the data collection techniques used by research papers in our dataset.

3.2 Research Question 2: Data Mining Techniques Used to Analyze and Predict Students' Performance

The main data mining approaches used in most of the analyzed studies are: (1) classification and (2) clustering. Table 8 summarizes the distribution of research articles in our dataset by the two data mining approaches described. The main data mining approach used

Table 6 Distribution of research articles by publication year

Year	Papers	Number of papers
2009	RP142	1
2010	RP28, RP35, RP113	3
2011	RP120, RP127, RP128	3
2012	RP23, RP25, RP33, RP43	4
2013	RP16, RP34, RP323	3
2014	RP9, RP44, RP67, RP123	4
2015	RP1, RP81, RP174, RP220	4
2016	RP10, RP12, RP22, RP36, RP75, RP198	6
2017	RP2, RP7, RP69, RP94	4
2018	RP5, RP277	2

Table 7 Summary of data collection techniques

Data collection technique	Description	Articles	No. of articles
e-Learning system logs	Logs obtained from e-Learning systems	RP1, RP7, RP9, RP23, RP25, RP33, RP34, RP36, RP44, RP94, RP113, RP198, RP220, RP277	14
Student information system data	Extracted data from student information systems, such as demographic data, admission data, grades, etc.	RP2, RP5, RP16, RP22, RP25, RP28, RP43, RP120, RP123, RP127, RP323	11
Surveys	General survey obtained from students directly	RP16, RP22, RP67, RP69, RP75, RP128, RP142	7
Course evaluations surveys	Answers of course evaluation surveys, generally obtained at the end of each course	RP10, RP12, RP25, RP35	4
Network access logs	Network dumb logs of students' activities on the internet and the university network	RP174	1

Table 8 Distribution of research articles by data mining approaches

Data mining approach	Papers	Number of papers
Classification	All	34
Clustering	RP2, RP23, RP25, RP34	4

is classification. It was found that all the research papers in the dataset have used the classification approach to classify and predict the students' performance. On the other hand, only 4 research papers have used the clustering along with classification which was useful in order to find out how many different groups of students available in the dataset and extract specific features of each group. This finding is in line with the findings of a prior study carried out by Peña-Ayala (2014), where it showed that the classification and clustering were the most typical data mining techniques used by EDM research.

Furthermore, we extracted 141 data mining techniques/algorithms used by the 34 papers in our dataset. Out of which, 74 were distinct. The algorithms are: 1NN, 3NN, ADTree, Apriori Algorithm, Artificial Neural Networks, BayesNet, Bivariate Regression, BP, C4.5 Decision Tree, CART Decision Tree, CHAID, CitationKNN, Clustering, CRT Decision Tree, Decision Tree (DT), DecisionStump, DTNB, EM, FarthestFirst, Feed-Forward Neural Network (FFNN), G3P-MI, GP-ICRM, Gradient Boosting (GBM), HierarchicalClusterer, IBk, ICRM v1, ICRM v2, ICRM v3, ICRM2, ID3 Decision Tree, J48, Jrip, K-Means Clustering, K-Nearest Neighbour (k-NN), LADTree, LGR, Locally weighted linear regression, Logistic Regression, MILR, Multi-layer Perceptron (MLP) neural network, Model Trees, Multi-logistic Regression (MLR), Naïve Bayes classifiers, NaiveBayesSimple, Neural Network (NN), NLPCA, Nnge, OneR, PART, Prism, Probabilistic ensemble SFAM classifier (PESFAM), Probabilistic Ensemble Simplified Fuzzy ARTMAP, Proportional Odd Model (POM), Radial Basis Function (RBF) Network, Random Forest, Random Tree, RBF Network, Regression, Regression neural network model (RNN), REPTree, Resonance Theory Mapping (PESFAM), Ridor, RIPPER, Rule Induction, sIB, SimpleCart, SimpleKMeans, SMO, Support Vector Machine (SVM), Support Vector Ordinal Regression (SVOR), System for Educational Data Mining (SEDM), Visualization, WINNOW, Xmeans. However, the most commonly used algorithms that were used in 4 or more research papers (i.e., in more than 10% of the papers), are shown in Table 9.

Furthermore, we merged similar algorithms together into one category, for example, ID3 and C4.5 are both decision trees, so we grouped them together under the decision tree category. After doing so, we ended up with 7 categories of algorithms. Table 10 shows the 7 groups of algorithms and the frequency of their usage in the analyzed research articles. As it can be observed from Table 10, the most commonly used categories of data mining algorithms are Decision Trees, Naïve Bayes classifiers, and Artificial Neural Networks.

4 Discussion

This study reports a systematic literature review regarding the students' academic performance in higher education. The study was designed to identify the most commonly studied factors that affect the students' performance, as well as, the most common data mining techniques applied to identify these factors. The study reviewed 34 research articles related

Table 9 Most commonly used data mining algorithms

Algorithm	Frequency	Percentage[a] (%)
Naïve Bayes classifiers	13	38.2
Support vector machine (SVM)	8	23.5
Logistic regression	6	17.6
K-Nearest neighbor (k-NN)	5	14.7
ID3 Decision tree	4	11.8
C4.5 Decision tree	4	11.8
Decision tree (DT)	4	11.8
Multi-layer perceptron (MLP) neural network	4	11.8
Neural network (NN)	4	11.8

[a]The percentage calculated is out of the total number of analyzed articles (N = 34)

Table 10 Most commonly used algorithms by category

Algorithm	Frequency	Percentage[a] (%)
Decision trees	35	24.8
Naïve Bayes classifiers	14	9.9
Artificial neural networks	13	9.2
Regression	12	8.5
Support vector machine	9	6.4
K-Nearest neighbor	8	5.7
K-Means	3	2.1
Other algorithms	47	33.3

[a]The percentage calculated is out of the total number of algorithms (N = 141)

to the subject and came up with results of research distribution across multiple dimensions. In terms of the first research direction, the main results showed that students' previous grades and class performance, students' e-Learning activity, students' demographics, and students' social information are generally the main factors affecting the students' academic performance in higher education.

According to the literature (Asif et al. 2017; Burgos et al. 2018; Gómez-Rey et al. 2016; Kotsiantis et al. 2010; Márquez-Vera et al. 2013, 2016), the first category which relates to the students' previous grades and class performance was also regarded as one of the influential factors that could affect the students' academic performance. This finding could be explained that students' performance remains equally the same across their educational life. In other words, if the student is in the habit of having good grades at the beginning of his/her studies, he/she would remain good for the rest of his/her academic life. This is equally the same for students who tend to score bad grades, they might also keep the same pattern across their study, and definitely, this could affect their performance in the current and future studies. For the higher educational institutions, these results could assist the education stakeholders to focus on the specific areas of weaknesses in the students' academic life and try to overcome these shortcomings by improving the students' educational outcomes and quality.

The results were also in agreement with the existing literature (Abdous et al. 2012; Burgos et al. 2018; Hung et al. 2012; Lara et al. 2014; Xing et al. 2015; Zafra and Ventura 2012), in which the second category of factors (i.e., students' e-Learning activity) was also found to have a significant impact on students' academic performance. This indicates that the more the students engage in e-Learning activities (e.g., accessing online material, solving online quizzes, and uploading assignments into the e-Learning system), the more likely the students achieve higher grades and improve their overall performance. Practically, these results could assist the educational institutions to concentrate on students' e-Learning activities and promote using e-Learning systems in order to increase the students' performance and education quality (Salloum et al. 2019).

Concerning the other two categories of factors which relate to the students' demographics and students' social information, it is posited that these categories are more students' specific categories that deal with the students' background and their surroundings behaviors. These two factors mainly depend on the students themselves, and that the students should take care of such factors and try to avoid any behaviors or social activities that might affect their academic performance. Moreover, the policy-makers can utilize such factors to create focus groups to take care of the students and provide them with special attention during their study in the institution.

With respect to the data collection techniques, the results pointed out that e-Learning system logs and student information system data were the most frequent data collection techniques used in the analyzed studies. Further research should consider other data collection techniques like surveys, course evaluation surveys, and network access logs as these techniques were less used in the existing literature.

In terms of the second research direction, the results pointed out that Decision Trees, Naïve Bayes classifiers, and Artificial Neural Networks are the most commonly used data mining algorithms. This finding comes in line with the findings of prior systematic literature review studies (Peña-Ayala 2014; Shahiri et al. 2015), where it was found that decision trees and Naïve Bayes classifiers are the most frequent data mining techniques used among EDM research. Given these results at hand, further research is suggested to refer to other data mining algorithms which might add more significant and reasonable conclusions.

Although the current systematic review partially shares the same research questions as with Shahiri et al. (2015) and Peña-Ayala (2014), it also has several differences that we need to discuss. First, the time span of the analyzed research articles ranges between 2002 and 2015 in terms of the review study conducted by Shahiri et al. (2015), and between 2010 and 2013 in terms of the review study carried out by Peña-Ayala (2014). In comparison with these reviews, the time span of the analyzed studies in this review study ranges between 2009 and 2018. Second, with regard to the search strategy, Shahiri et al. (2015) has mainly focused on collecting studies related to "students' performance" and "educational data mining", whereas Peña-Ayala (2014) has concentrated on collecting studies related to "EDM approaches" and "EDM tools". To make this review study more distinctive, it has focused on collecting studies related to "factors affecting students' performance" and "data mining or educational data mining techniques". Changing the search strategy allows us to retrieve different articles from those that were retrieved in previous reviews.

Third, Shahiri et al. (2015) concluded that Decision tree and Neural Networks were the most commonly used data mining techniques, whereas Peña-Ayala (2014) indicated that Bayes theorem, Decision tree, instance-based learning (IBL), and hidden Markov model (HMM) were the most frequent techniques used. In this study, the results were partially supported, in which Decision tree, Naïve Bayes classifiers, and Artificial Neural Networks were observed to be the most common data mining techniques used to predict students'

performance. In general, the differences between this review study and the previous ones lie in the gap between time span, search strategy, and results achieved. Thus, this study could serve as a comprehensive reference for pursuing further research in EDM in general, and students' performance in particular.

5 Conclusion and Future Work

In this paper, we identified the most common and widely studied factors that affect students' performance in higher education, as well as, the most common data mining approaches, techniques, and algorithms used to classify and predict students' performance. We followed a systematic literature review methodology that consisted of multiple phases and steps. This process is started by planning the review, from forming up the research questions, through setting up the inclusion and exclusion criteria, until deciding the data extraction strategy. Furthermore, the second phase consisted of the steps for conducting the review, kicked off by searching and identifying the research papers for the literature review, passing by assessing the quality of the selected research papers, and ended by extracting the data and synthesize it. Finally, we ended up by reporting the results of our SLR research, which concluded that the most common and widely used factors for predicting students' performance in higher education are students' previous grades and class performance, students' e-Learning activity, students' demographics, and students' social information. Additionally, the results also showed that the most common and widely used data mining techniques in the EDM field are Decision Trees, Naïve Bayes classifiers, and Artificial Neural Networks.

As a future work, researchers can benefit from the outcomes of this systematic literature review by employing it to their future research, particularly, the main results that highlight the most frequently used categories of factors affecting students' performance, as well as, the most frequently used data mining techniques. Not to mention, having a generic set of factors' categories provides several possibilities to tailor the use of these categories and come up with specific factors within the category of each educational institution, since it might differ from one place to another, and from time to time. As a limitation, this study has focused on some databases in terms of articles collection. In addition, future attempts may consider other databases and search engines for articles collection in order to leverage the number of analyzed studies.

Appendix

See Tables 11, 12, 13 and 14.

Table 11 Selected research papers

Paper ID	Source	Journal	Author
RP1	ScienceDirect	Computers in Human Behavior	Xing et al. (2015)
RP2	ScienceDirect	Computers and Education	Asif et al. (2017)
RP5	ScienceDirect	Journal of Business Research	Fernandes et al. (2018)
RP7	ScienceDirect	Computers and Electrical Engineering	Burgos et al. (2018)
RP9	ScienceDirect	Computers and Education	Lara et al. (2014)
RP10	EBSCO	Expert Systems	Gómez-Rey et al. (2016)
RP12	ProQuest	Informatics in Education	Jiang et al. (2016)
RP16	ProQuest	Applied Intelligence	Márquez-Vera et al. (2013)
RP22	EBSCO	Expert Systems	Márquez-Vera et al. (2016)
RP23	JSTOR	Journal of Educational Technology and Society	Abdous et al. (2012)
RP25	JSTOR	Journal of Educational Technology and Society	Hung et al. (2012)
RP28	ScienceDirect	Knowledge-Based Systems	Kotsiantis et al. (2010)
RP33	ScienceDirect	Applied Soft Computing	Zafra and Ventura (2012)
RP34	ScienceDirect	Computers and Education	Romero et al. (2013)
RP35	EBSCO	Applied Stochastic Models in Business and Industry	Costantini et al. (2010)
RP36	EBSCO	Expert Systems	Gamulin et al. (2016)
RP43	ProQuest	The Artificial Intelligence Review	Kotsiantis (2012)
RP44	ScienceDirect	Computers in Human Behavior	Hu et al. (2014)
RP67	JSTOR	European Journal of Open, Distance and E-Learning	Yukselturk et al. (2014)
RP69	JSTOR	Journal of Computer Science	Abazeed and Khder (2017)
RP75	Google Scholar	International Journal of Advanced Computer Science and Applications	Abu Saa (2016)
RP81	JSTOR	Indian Journal of Science and Technology	Anuradha Bharathiar (2015)
RP94	JSTOR	The Electronic Journal of Information Systems in Developing Countries	Mwalumbwe and Mtebe (2017)
RP113	ScienceDirect	Computers and Education	Macfadyen and Dawson (2010)
RP120	Google Scholar	International Journal of Advanced Computer Science and Applications	Baradwaj and Pal (2012)

Table 11 (continued)

Paper ID	Source	Journal	Author
RP123	Google Scholar	World Journal of Computer Application and Technology	Badr El Din Ahmed and Sayed Elaraby (2014)
RP127	Google Scholar	International Journal of Computer Science and Information Technologies	Pandey and Pal (2011)
RP128	Google Scholar	International Journal of Computer Science and Information Security	Bhardwaj and Pal (2012)
RP130	Google Scholar	International Journal of Innovative Technology and Creative Engineering	Yadav et al. (2012)
RP136	Google Scholar	World of Computer Science and Information Technology Journal	Yadav and Pal (2012)
RP142	ScienceDirect	Computers and Education	Araque et al. (2009)
RP174	ProQuest	International Conference on Digital Information and Communication Technology and its Applications	Zhou et al. (2015)
RP198	ScienceDirect	Computers and Education	Cerezo et al. (2016)
RP220	ProQuest	The International Journal of Information and Learning Technology	Chamizo-Gonzalez et al. (2015)
RP277	EBSCO	Journal of AI and Data Mining	Hasheminejad and Sarvmili (2018)
RP323	ScienceDirect	Computers and Education	Huang and Fang (2013)

Table 12 Quality scores and percentages

Paper ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Quality score	Percentage (%)
RP10	1	1	1	1	1	0	1	1	1	8	88.89
RP22	1	1	1	1	1	0	1	1	1	8	88.89
RP113	1	1	1	1	1	0	0.5	1	1	7.5	83.33
RP1	1	1	1	1	1	0	0.5	0.5	1	7	77.78
RP2	1	1	1	1	1	0	0.5	0.5	1	7	77.78
RP5	1	1	1	1	1	0	0.5	0.5	1	7	77.78
RP7	1	1	1	1	1	0	0.5	0.5	1	7	77.78
RP9	1	1	1	1	1	0	0.5	0.5	1	7	77.78
RP12	1	1	1	1	1	0	0	1	1	7	77.78
RP23	1	1	1	1	1	0	0.5	1	0.5	7	77.78
RP28	1	1	1	1	1	0	0.5	0.5	1	7	77.78
RP33	1	1	1	1	1	0	0.5	0.5	1	7	77.78
RP34	1	1	1	1	1	0	0.5	0.5	1	7	77.78
RP44	1	1	1	1	1	0	0.5	0.5	1	7	77.78
RP75	1	1	1	1	1	0	0	1	1	7	77.78
RP142	1	1	1	1	1	0	0.5	0.5	1	7	77.78
RP174	1	1	1	1	1	0	0	1	1	7	77.78
RP198	1	1	1	1	1	0	0.5	0.5	1	7	77.78
RP323	1	1	1	1	1	0	0.5	0.5	1	7	77.78
RP220	1	1	1	1	0.5	0	0.5	0.5	1	6.5	72.22
RP16	1	1	1	1	0.5	0	0	0.5	1	6	66.67
RP25	1	1	1	1	0.5	0	0.5	0.5	0.5	6	66.67
RP123	1	0.5	0.5	1	1	0	0	1	1	6	66.67
RP35	1	1	1	1	0.5	0	0	0.5	0.5	5.5	61.11
RP36	1	1	1	1	0.5	0	0.5	0	0.5	5.5	61.11
RP43	1	1	1	1	0.5	0	0	0.5	0.5	5.5	61.11
RP67	1	1	1	0.5	0.5	0	0.5	0	1	5.5	61.11
RP69	1	1	1	0.5	1	0	0	0	1	5.5	61.11
RP94	1	0.5	0.5	1	0.5	0	1	0	1	5.5	61.11
RP127	1	0.5	0.5	1	1	0	0	0.5	1	5.5	61.11
RP277	1	1	1	1	0.5	0	0	0.5	0.5	5.5	61.11
RP120	1	0.5	0.5	0.5	0	0	1	1	0.5	5	55.56
RP81	1	0	0.5	1	0.5	0	0.5	0.5	0.5	4.5	50.00
RP128	1	0.5	0.5	0.5	1	0	0	0.5	0.5	4.5	50.00
RP130	1	0.5	0.5	0.5	0.5	0	0	0.5	0.5	4	44.44
RP136	1	0.5	0.5	0.5	0.5	0	0	0.5	0.5	4	44.44

Table 13 Summary of factors influencing students' performance and used data mining approaches and techniques

Paper ID	Factors category	Factors found significant	Data mining approach	Data type(s)
RP1	Students e-Learning activity	Chat logs of all messages that students send to each other in the group Awareness records of actions of erasing the chat messages on the chat bar Geogebra logs of information on how students virtually construct a geometry artifact (adding a point or updating a segment) System logs of students joining a virtual room, leaves a virtual room or views different tabs WhiteBoard logs of more specific actions on how tools are being used in the white board areas such as resizing objects or creating a textbox	Classification	e-Learning system logs
RP2	Students previous grades and class performance	High school marks (total and subject specific) First and second year university courses' marks	1. Classification 2. Clustering	Admission data
RP5	1. Students environment 2. Students demographics 3. Students previous grades and class performance	Grades for the first 2 months Student's place of residence—neighborhood School name School subjects Absences Student's place of residence—city Age	Classification	Student information system data
RP7	Students e-Learning activity	12 Assessment activities from e-Learning system Teaching schedule	Classification	e-Learning system logs
RP9	Students e-Learning activity	Number of virtual classroom accesses by the student in the week in question Number of different days of the week on which the student accesses the virtual classroom Whether or not the resource has been visualized in the week in question Number of times that the student has visualized the resource in the week in question	Classification	e-Learning system logs

Table 13 (continued)

Paper ID	Factors category	Factors found significant	Data mining approach	Data type(s)
RP10	1. Instructor attributes 2. Students previous grades and class performance 3. Course attributes	Instructor's knowledge Instructor's effective use of the class hours Instructor's coherence with lesson plan Openness and respect of the instructor to students' views Instructor's positive approach to students Instructor readiness for classes Instructor explanations about the course and instructor helpfulness	Classification	Course evaluations surveys
RP12	1. Instructor attributes 2. Course attributes	Instructor's organization and clarity Instructor's response to questions Instructor's visual presentation Instructor's encouragement to think independently Instructor's attitude towards teaching Professor-class relationship Difficulty of concepts covered Contribution of assignments to understanding of concepts How well tests reflect the course material Attendance (the number of evaluations received divided by course enrolment)	Classification	Course evaluations surveys

Table 13 (continued)

Paper ID	Factors category	Factors found significant	Data mining approach	Data type(s)
RP16	1. Students previous grades and class performance 2. Students demographics 3. Students social data	Scores in specific subjects Level of motivation GPA in secondary school Age Number of brothers/sisters Classroom/group Smoking habits Studying in group Marital status Time spent doing exercises	Classification	1. Surveys 2. Student information system data
RP22	1. Students previous grades and class performance 2. Students demographics 3. Students social data	GPA in secondary school Classroom/group enrolled Number of students in the group/class Age Attendance during morning/evening sessions Having a job Mother's level of education	Classification	1. Surveys 2. Student information system data
RP23	Students e-Learning activity	Students activity data from an online video e-learning system Number of questions Number of chat messages Total login times Final grade	1. Classification 2. Clustering	e-Learning system logs

Table 13 (continued)

Paper ID	Factors category	Factors found significant	Data mining approach	Data type(s)
RP25	1. Students e-learning activity 2. Students demographics 3. Course evaluations	1. Students e-Learning activity Average frequency of logins per course Average frequency of tab accessed per course Average frequency of module accessed per course Average frequency of clicks per course Average frequency of course accessed per course Average frequency of page accessed per course Average frequency of course content accessed per course Average number of discussion board entries per course 2. Students demographics: age, gender, graduation year, city, school district, number of online course(s) taken, number of online course(s) passed, number of online course(s) failed, and final grade average 3. Student information: number of courses taken, number of courses failed, number of courses passed, average individual student pass rate for all courses in academic year 2009–2010	1. Classification 2. Clustering	1. e-Learning system logs 2. Student information system data 3. Course evaluations surveys
RP28	Students previous grades and class performance	1st written assignment 2nd written assignment 3rd written assignment 4th written assignment	Classification	Student information system data

Table 13 (continued)

Paper ID	Factors category	Factors found significant	Data mining approach	Data type(s)
RP33	Students e-Learning activity	Number of pieces of coursework done by the user in the course Total time in seconds that the user has taken in the assignment section Number of messages sent by the user in the forum Number of messages read by the user in the forum Total time in seconds that the user has taken in the forum section Number of quizzes seen by the user Number of quizzes passed by the user Number of quizzes failed by the user Total time in seconds that the user has taken in the quiz section	Classification	e-Learning system logs
RP34	Students e-Learning activity	Number of messages written by the student Number of words written by the student Average score on the instructor's evaluation of the student's messages Degree centrality of the student Degree prestige of the student	1. Classification 2. Clustering	e-Learning system logs

Table 13 (continued)

Paper ID	Factors category	Factors found significant	Data mining approach	Data type(s)
RP35	Course evaluations	Program workload Program organization of teaching Keep scheduled hours Clear exam rules Availability of lecturer outside class Student's previous knowledge of the topic Teacher ability to motivate Clarity of teaching Availability of lecturer inside class On schedule with program Workload-credit ratio Prescribed reading list Adequacy of lecture hall Student interest in topic Overall class satisfaction	Classification	Course evaluations surveys
RP36	Students e-Learning activity	Student access time series Number of clicks per course.	Classification	e-Learning system logs
RP43	1. Students previous grades and class performance 2. Students demographics	4th Written assignment 3rd Written assignment	Classification	Student information system data

Table 13 (continued)

Paper ID	Factors category	Factors found significant	Data mining approach	Data type(s)
RP44	Students e-Learning activity	Total time online (s) Number of course material viewed (by material category) (s) Average time per session (s) Total time material viewed (s) Number of course material viewed # Course material viewed (by material category)/ number of Course material released to date Number of logins Average time material viewed (s) Total time course material viewed (by material category) (s)	Classification	e-Learning system logs
RP67	1. Students demographics 2. Students experience data	Online learning readiness Previous online experience Gender Online technologies self-efficacy Age Prior knowledge	Classification	Surveys
RP69	1. Students demographics 2. Students previous grades and class performance	Gender High school grade Major in high school Previous GPA Number of courses registered Sponsor Advisory visit English score Attendance Core versus elective Study time Performance	Classification	Surveys

Table 13 (continued)

Paper ID	Factors category	Factors found significant	Data mining approach	Data type(s)
RP75	1. Students demographics 2. Students previous grades and class performance 3. Students social data	Gender High school grade Mother occupation status Discount	Classification	Surveys
RP81	1. Students demographics 2. Students previous grades and class performance 3. Students social data	Previous semester marks Family annual income Student category Family size Attendance High school grade Assignment performance	Classification	Not reported
RP94	Students e-Learning activity	Interactions with peers Number of exercises performed Number of forum posts	Classification	e-Learning system logs
RP113	Students e-Learning activity	Total # discussion messages posted Total number of online sessions Total time online # Files viewed # Assessments finished # Assessments started # Reply discussion messages posted # Mail messages sent # Assignments submitted # Discussion messages read # Web links viewed # New discussion messages posted # Mail messages read	Classification	e-Learning system logs

Table 13 (continued)

Paper ID	Factors category	Factors found significant	Data mining approach	Data type(s)
RP120	Students previous grades and class performance	Previous semester marks Class test grade Attendance Assignment Lab work	Classification	Student information system data
RP123	1. Students demographics 2. Students previous grades and class performance	Midterm marks Lab test grades Students practice Homework Seminar performance High school branch Attendance	Classification	Student information system data
RP127	Students demographics	Gender Language medium Stream of bachelor's degree Division obtained	Classification	Student information system data
RP128	1. Students demographics 2. Students previous grades and class performance 3. Students social data	Students grade in senior secondary education Living location Medium of teaching Mother's qualification Students other habit Family annual income status Students family status	Classification	Surveys
RP142	1. Students demographics 2. Students previous grades and class performance 3. Students social data	Father's education level Mother's education level Academic performance rate Average round Mode round Access year	Classification	Surveys

Table 13 (continued)

Paper ID	Factors category	Factors found significant	Data mining approach	Data type(s)
RP174	Students social data	Number of records on each category of websites Duration of watching online videos Students' grades on advanced mathematics	Classification	Network access logs
RP198	Students e-Learning activity	Total time spent on practical tasks The time taken to hand in the task since the task was made available in the LE Number of words in forum posts	Clustering	e-Learning system logs
RP220	Students e-Learning activity	Assignment upload Forum add post Forum update post Forum view discussion Assignment view Assignment view all Course view Forum view forum Resource view	Classification	e-Learning system logs
RP277	Students e-Learning activity	Course identification number Number of assignments done Number of quizzes passed Number of quizzes failed Number of messages send to forum Number of messages read on the forums Total time used on assignments Total time used on quizzes Total time used on forum	Classification	e-Learning system logs

Table 13 (continued)

Paper ID	Factors category	Factors found significant	Data mining approach	Data type(s)
RP323	Students previous grades and class performance	Cumulative GPA Statistics grade Calculus I grade Calculus II grade Physics grade Dynamics mid-exam #1 score Dynamics mid-exam #2 score Dynamics mid-exam #3 score	Classification	Student information system data

Table 14 Description of factors' categories (extended)

Category	Papers	Number of articles
Students e-Learning activity	RP1, RP7, RP9, RP23, RP25, RP33, RP34, RP36, RP44, RP94, RP113, RP198, RP220, RP277	14
Students Previous grades and class performance	RP2, RP5, RP10, RP16, RP22, RP28, RP43, RP69, RP75, RP81, RP120, RP123, RP128, RP142, RP323	15
Students environment	RP5	1
Students demographics	RP5, RP16, RP22, RP25, RP43, RP67, RP69, RP75, RP81, RP123, RP127, RP128, RP142	13
Instructor attributes	RP10, RP12	2
Course attributes	RP10, RP12	2
Students social information	RP16, RP22, RP75, RP81, RP128, RP142, RP174	7
Course evaluations	RP25, RP35	2
Students experience information	RP67	1

References

Abazeed, A., & Khder, M. (2017). A classification and prediction model for student's performance in university level. *Journal of Computer Science, 13,* 228–233.

Abdous, M., He, W., & Yen, C. J. (2012). Using data mining for predicting relationships between online question theme and final grade. *Educational Technology and Society, 15*(3), 77–88.

Abu Saa, A. (2016). Educational data mining and students' performance prediction. *Internation Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/IJACSA.2016.070531.

Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Mining student information system records to predict students' academic performance. In *International conference on advanced machine learning technologies and applications* (pp. 229–239). Berlin: Springer.

Al-Emran, M., Mezhuyev, V., Kamaludin, A., & Shaalan, K. (2018). The impact of knowledge management processes on information systems: A systematic review. *International Journal of Information Management, 43,* 173–187.

Al-Qaysi, N., Mohamad-Nordin, N., & Al-Emran, M. (2018). A systematic review of social media acceptance from the perspective of educational and information systems theories and models. *Journal of Educational Computing Research*. https://doi.org/10.1177/0735633118817879.

Anuradha Bharathiar, C., & Velmurugan, T. (2015). A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. *Indian Journal of Science and Technology*. https://doi.org/10.17485/ijst/2015/v8i.

Araque, F., Rolddsán, C., & Salguero, A. (2009). Factors influencing university drop out rates. *Computers and Education*. https://doi.org/10.1016/j.compedu.2009.03.013.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers and Education*. https://doi.org/10.1016/j.compedu.2017.05.007.

Badr El Din Ahmed, A., Sayed Elaraby, I., & Sayed Elaraby, I. (2014). Data mining: A prediction for student's performance using classification method. *World Journal of Computer Application and Technology*. https://doi.org/10.13189/wjcat.2014.020203.

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*. https://doi.org/10.1007/s10639-017-9616-z.

Baradwaj, B., & Pal, S. (2012). Mining educational data to analyze student's performance. *International Journal of Advanced Computer Science and Applications, 2*(6), 63–69.

Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-014-9223-7.

Bhardwaj, B. K., & Pal, S. (2012). Data mining: A prediction for performance improvement using classification. *(IJCSIS) International Journal of Computer Science and Information Security, 9*(4), 1–5.

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering*. https://doi.org/10.1016/j.compeleceng.2017.03.005.

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers and Education*. https://doi.org/10.1016/j.compedu.2016.02.006.

Chamizo-Gonzalez, J., Cano-Montero, E. I., Urquia-Grande, E., & Muñoz-Colomina, C. I. (2015). Educational data mining for improving learning outcomes in teaching accounting within higher education. *International Journal of Information and Learning Technology*. https://doi.org/10.1108/IJILT-08-2015-0020.

Costantini, P., Linting, M., & Porzio, G. C. (2010). Mining performance data through nonlinear PCA with optimal scaling. *Applied Stochastic Models in Business and Industry*. https://doi.org/10.1002/asmb.771.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2018). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2018.02.012.

Gamulin, J., Gamulin, O., & Kermek, D. (2016). Using Fourier coefficients in time series analysis for student performance prediction in blended learning environments. *Expert Systems*. https://doi.org/10.1111/exsy.12142.

Gómez-Rey, P., Fernández-Navarro, F., & Barberà, E. (2016). Ordinal regression by a gravitational model in the field of educational data mining. *Expert Systems*. https://doi.org/10.1111/exsy.12138.

Hasheminejad, S. M., & Sarvmili, M. (2018). S3PSO: Students' performance prediction based on particle swarm optimization. *Journal of AI and Data Mining, 7,* 77–96.

Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2014.04.002.

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers and Education*. https://doi.org/10.1016/j.compedu.2012.08.015.

Hung, J., Hsu, Y.-C., & Rice, K. (2012). Integrating data mining in program evaluation of K-12 online education. *Educational Technology and Society*. https://doi.org/10.1207/s15327752jpa8502.

Jiang, Y. H., Javaad, S. S., & Golab, L. (2016). Data mining of undergraduate course evaluations. *Informatics in Education*. https://doi.org/10.15388/infedu.2016.05.

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (pp. 1–57). Software Engineering Group, School of Computer Science and Mathematics, Keele University.

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering: A systematic literature review. *Information and Software Technology*. https://doi.org/10.1016/j.infsof.2008.09.009.

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-011-9234-x.

Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2010.03.010.

Lara, J. A., Lizcano, D., Martínez, M. A., Pazos, J., & Riera, T. (2014). A system for knowledge discovery in e-Learning environments within the European higher education area: Application to student data from Open University of Madrid, UDIMA. *Computers and Education, 72,* 23–36. https://doi.org/10.1016/j.compedu.2013.10.009.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J., et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2009.06.006.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education*. https://doi.org/10.1016/j.compedu.2009.09.008.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*. https://doi.org/10.1111/exsy.12135.

Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence*. https://doi.org/10.1007/s10489-012-0374-8.

Mwalumbwe, I., & Mtebe, J. S. (2017). Using learning analytics to predict students' performance in moodle learning management system: A case of Mbeya University of science and technology. *Electronic Journal of Information Systems in Developing Countries*. https://doi.org/10.1002/j.1681-4835.2017.tb00577.x.

Pandey, U. K., & Pal, S. (2011). Data mining: A prediction of performer or underperformer using classification. *International Journal of Computer Science and Information Technologies, 2*, 686–690.

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.

Romero, C., López, M. A., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers and Education*. https://doi.org/10.1016/j.compedu.2013.06.009.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2006.04.005.

Salloum, S. A., Al-Emran, M., Shaalan, K., & Tarhini, A. (2019). Factors affecting the E-learning acceptance: A case study from UAE. *Education and Information Technologies, 24*(1), 509–530. https://doi.org/10.1007/s10639-018-9786-3.

Shahiri, A. M., Husain, W., & 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.

Wook, M., Yusof, Z. M., & Nazri, M. Z. A. (2017). Educational data mining acceptance among undergraduate students. *Education and Information Technologies*. https://doi.org/10.1007/s10639-016-9485-x.

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2014.09.034.

Yadav, S., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. *International Journal of Innovative Technology and Creative Engineering, 1*, 13–19.

Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. *World of Computer Science and Information Technology Journal WCSIT*. https://doi.org/10.1142/9789812771728_0012.

Yukselturk, E., Ozekes, S., Türel, Y. K., Education, C., Ozekes, S., Türel, Y. K., et al. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and E-Learning*. https://doi.org/10.2478/eurodl-2014-0008.

Zafra, A., & Ventura, S. (2012). Multi-instance genetic programming for predicting student performance in web based educational environments. *Applied Soft Computing*. https://doi.org/10.1016/j.asoc.2012.03.054.

Zhou, Q., Zheng, Y., & Mou, C. (2015). Predicting students' performance of an offline course from their online behaviors. In *2015 5th international conference on digital information and communication technology and its applications, DICTAP 2015*. https://doi.org/10.1109/DICTAP.2015.7113173.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.