Psicothema

Psicothema 2021, Vol. 33, No. 3, 456-465
doi: 10.7334/psicothema2021.62

ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2021 Psicothema
www.psicothema.com

Early Prediction of Student Learning Performance Through Data Mining: A Systematic Review

Javier López-Zambrano[1], Juan Alfonso Lara Torralbo[2], and Cristóbal Romero[3]
[1] Escuela Superior Politécnica Agropecuaria de Manabí, [2] Madrid Open University, and University of Córdoba[3]

Abstract

Background: Early prediction of students' learning performance using data mining techniques is an important topic these days. The purpose of this literature review is to provide an overview of the current state of research in that area. **Method:** We conducted a literature review following a two-step procedure, looking for papers using the major search engines and selection based on certain criteria. **Results:** The document search process yielded 133 results, 82 of which were selected in order to answer some essential research questions in the area. The selected papers were grouped and described by the type of educational systems, the data mining techniques applied, the variables or features used, and how early accurate prediction was possible. **Conclusions:** Most of the papers analyzed were about online learning systems and traditional face-to-face learning in secondary and tertiary education; the most commonly-used predictive algorithms were J48, Random Forest, SVM, and Naive Bayes (classification), and logistic and linear regression (regression). The most important factors in early prediction were related to student assessment and data obtained from student interaction with Learning Management Systems. Finally, how early it was possible to make predictions depended on the type of educational system.

Keywords: Educational Data Mining; Learning Analytics; Early prediction of academic performance; Early Warning Systems; Detection of students at-risk of Dropping-out.

Resumen

Predicción Temprana del Rendimiento Académico con Minería de Datos: una Revisión Sistemática. **Antecedentes:** la predicción temprana del rendimiento académico mediante técnicas de minería de datos es un campo de estudio emergente, que se pretende analizar por medio de este artículo de revisión. **Método:** se ha revisado la literatura existente por medio de un proceso de búsqueda de artículos en los principales motores de búsqueda, y de selección de los mismos de acuerdo con ciertos criterios. **Resultados:** el proceso de búsqueda reportó 133 resultados, de los cuales 82 fueron seleccionados para dar respuesta a las preguntas de investigación planteadas. Se han agrupado los trabajos encontrados para poder dar respuesta a las preguntas por tipo de sistema educativo, técnicas de minería de datos aplicadas, variables empleadas y grado de anticipación con el que se puede predecir. **Conclusiones:** la mayor parte de los trabajos publicados corresponden a sistemas de aprendizaje en línea y presenciales-tradicionales en educación secundaria y terciaria; los algoritmos más utilizados el J48, Random Forest, SVM, Naive Bayes (clasificación), y la regresión logística y lineal (regresión); los datos de evaluación y los obtenidos de la interacción del estudiante con el entorno de aprendizaje son las variables más relevantes; finalmente, la anticipación en la predicción varía según el tipo de sistema educativo.

Palabras clave: Data Mining Educativo; Analítica de Aprendizaje; predicción temprana del rendimiento académico; sistemas de detección temprana; estudiantes en riesgo de abandono.

Predicting studentsí learning performance is a challenging but essential task in education (Romero & Ventura, 2013). The prediction of academic performance is important not only to help students take control of their own learning and become self-regulated learners but also to allow educators to identify at-risk students and reduce the chances of failure (Bogarín et al., 2018). This is a difficult task because of the many possible factors that can influence student performance. In order to shed some light on this problem, EDM (Educational Data Mining) and Learning Analytics (LA) techniques have been successfully applied, mainly in e-learning environments (LMS -Learning Management Systems-, MOOC -Massive Open Online Courses-; etc.), where the volume of generated data is especially large and the studentsí activity reflects their learning processes (Castro et al., 2007). Data with information about students can also be gathered from traditional face-to-face education environments and from blended learning environments (B-learning).

The use of EDM and LA techniques to analyze these large amounts of data has produced interesting, interpretable, useful and novel information about learners (Fayyad et al., 1996). The application of Data Mining (DM) techniques to information about learning activities produced in educational environments is known as EDM (Barnes et al., 2009). EDM uses the same DM techniques with certain adaptations depending on the specific problems to be solved (Romero & Ventura, 2020). One of its main tasks is to predict student learning performance (failure, success, school dropout, etc.). LA can be defined as the measurement, collection,

Received: February 19, 2021 • Accepted: May 20, 2021
Corresponding author: Juan Alfonso Lara Torralbo
Department of Computer Science
Madrid Open University (UDIMA)
28400 Collado Villalba (Spain)
e-mail: juanalfonso.lara@udima.es

analysis, and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs (Siemens, 2013). Hence, EDM and LA are deeply related fields, and share the common objective of predicting and guiding student learning.

Early prediction can be defined as the application of predictive models that use key variables to accurately predict student failure or dropout as early as possible (Berens et al., 2018; Yu et al., 2018). It also refers to the technological information in the management of studentsí academic work for the early detection of their potential or real academic problems (Wang et al., 2018). It is necessary to detect at-risk students as early as possible and thus provide early intervention or care to help students succeed and to prevent them from quitting or failing. A wide range of student information can be used to make early predictions of student performance. Examples include student-completed questionnaires (Krotseng, 1992), lessons and activities in the early stages of courses (Costa et al., 2017), student performance and demographic data (Berens et al., 2018), activities and comments on evaluations to analyze feelings (Yu et al., 2018), records from online environments (Howard et al., 2018), and affective and emotive variables (Mújica et al., 2019) among others.

Early prediction is a challenging task for the EDM field due to the many factors that can influence a studentís final status. It is a critical issue in education because it concerns many students at all stages (primary education, secondary education, and tertiary or higher education) and in schools and universities all over the world. Early prediction is also essential in order to identify at-risk students as early as possible in order to implement programs that provide appropriate, effective prevention strategies, give advice or recommendations, and carry out remedial actions or interventions (Romero & Ventura, 2019).

Although there are some review papers about the prediction of academic performance (Ameen et al., 2019; Felix et al., 2018), the identification of at-risk students in general (Nik Nurul Hafzan et al., 2019), the use of exclusively LMS course data for prediction (Na & Tasir, 2018), and the application of Early Warning Systems óEWSó (McMahon & Sembiante, 2020) (Liz-Domínguez et al., 2019), none of them focus on early prediction through data mining techniques. This is the main reason that the current survey is necessary.

In this paper, rather than only analyzing studies about early prediction, an analysis was also carried out looking at different aspects related to early prediction, such as the education systems considered, the most commonly-used techniques and algorithms, how early it is possible to predict, and which are the most commonly-used variables or attributes.

The purpose of this survey is to conduct a systematic review of the literature about early prediction of academic performance in order provide readers with an introduction to the application of EDM/LA for early prediction and thus answer the following research questions: In what type of educational system has early prediction been applied most often? What techniques have been used most often? Which specific algorithms are the most used, and which have produced the best prediction results? How early can academic performance be predicted with acceptable accuracy? What specific variables or attributes have been used and demonstrated better performance?

The major original scientific contributions of this paper are:

- We present and summarize the most important scientific literature about the use of data mining techniques for early prediction of student performance.
- We have taxonomized those references and grouped them by the type of educational system.
- We have discovered and presented a series of research niches and opportunities in the area by analyzing aspects such as the most-used techniques, the attributes used, and how early the predictions of academic performance can be made.

This paper is organized as follows: The procedure section describes the process used for the systematic review. The results and discussion sections describe the studies selected, and the answers to the five research questions. Finally, the conclusions and future lines of research are presented.

Method

Procedure

Search strategy

We followed the systematic literature review procedure by Tranfield et al. (2003). Systematic reviews begin by defining a review protocol that specifies the research questions and the methods that will be used to perform the review. Following that, we defined the keywords and the explicit inclusion and exclusion criteria for searching for and selecting papers about early prediction. A double filter process was applied to discard papers that did not meet the inclusion criteria after reading the abstract (first filter) and the full paper (second filter).

We used Google Scholar, Web of Science, and Scopus search engines in order to search for all academic papers about early prediction published up to November 2020. The search used the following search terms:

1. “Early prediction” AND “Data Mining” AND (“academic performance” OR “at-risk students” OR dropouts)
2. “Early prediction” AND “Learning Analytics” AND (“academic performance” OR “at-risk students” OR dropouts)
3. “Early detection” AND “Data Mining” AND (“academic performance” OR “at-risk students” OR dropouts)
4. “Early detection” AND “Learning Analytics” AND (“academic performance” OR “at-risk students “OR dropouts)
5. “Early warning systems” AND (“academic performance” OR “at-risk students” OR dropouts)

Selecting papers

The papers were selected by reading both the abstract and full content of the papers initially downloaded from the search and applying the following inclusion and exclusion rules:

- Inclusion: articles focused exclusively on the topic of early prediction of student performance through EDM techniques.
- Exclusion: articles that did not actually perform early prediction of students’ performance through EDM techniques despite containing some of the search keywords.

Results

Starting from the search using the keywords noted above, a total of 133 papers were downloaded. There were 97 journal articles, 29 articles from international conferences, and 7 items corresponding to types such as books, reports, and doctoral theses.

As Figure 1 shows, the preliminary search identified 133 papers published up to November 2020 whose titles included the defined keywords. The abstract of each paper was read, leading to 17 papers being discarded for not doing early prediction. The remaining 116 papers were read in full, and 34 additional papers were discarded for the same reason. Many papers contained early prediction in the titles, but in reality they described classical prediction by using all the information provided at the end of the courses. The remaining 82 papers were used to answer the five research questions.

After reading the final selection of 82 articles, an analysis was carried out from various perspectives in order to answer each of the 5 research questions. In the sections, we describe and discuss the results and give an overview of the literature about the topic.

Discussion

Figure 2 shows that the first papers were published in the 1990s, which indicates that early prediction is not a new concern. However, it was not until 2008 when further research in this regard began,

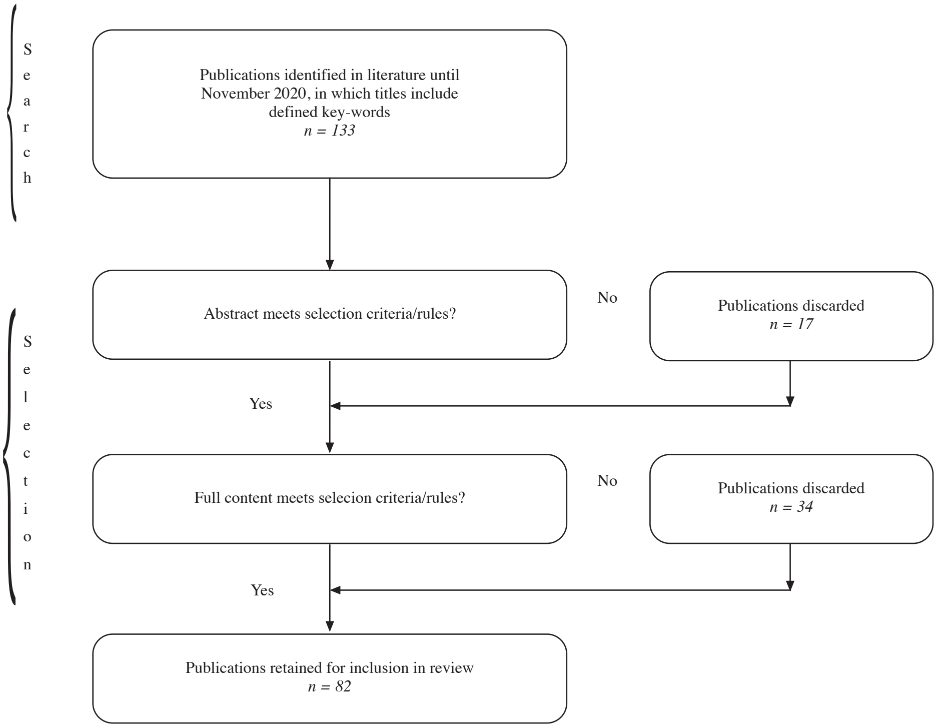
Search
Publications identified in literature until November 2020, in which titles include defined key-words
n = 133
Selection
Abstract meets selection criteria/rules?
No
Publications discarded
n = 17
Yes
Full content meets selecion criteria/rules?
No
Publications discarded
n = 34
Yes
Publications retained for inclusion in review
n = 82

Figure 1. Procedure flowchart

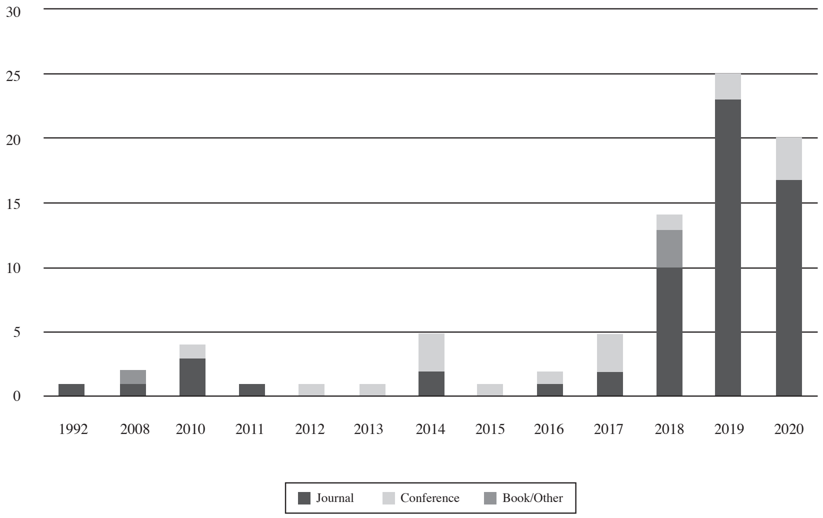

Figure 2. Number of papers published per year

and the most significant contributions came in the last decade. In addition, we have noticed that in the last 4 years (2017-2020) there have been a significant number of contributions.

Table 1 shows the 5 most-cited papers about early prediction of student learning performance. The first ranked paper affirms that LMS-generated student data can be used for identifying at-risk students and can allow more timely pedagogical interventions (Macfadyen & Dawson, 2010). The second describes the goals and objectives of the Open Academic Analytics Initiative (OAAI), and describes the process and challenges of collecting, organizing and mining student data to predict academic risk and the results of interventions with at□risk students (Jayaprakash et al., 2014). The third paper explores the socio-demographic variables and study environment that may influence student persistence or dropout and examines the extent to which these factors help us in pre-identifying successful and unsuccessful students (Kovačić, 2010). The fourth paper seeks to identify significant behavioral indicators of learning using LMS data regarding online course achievement (You, 2016). The fifth paper in the ranking presents a comparative study on the effectiveness of educational data mining techniques for early prediction of students likely to fail in introductory programming courses (Costa et al., 2017).

What type of educational system has early prediction been applied to most often?

Early prediction can be applied to various types of educational systems and levels. These include: Traditional education, referring to long-established practices traditionally used in schools (in-person); E-learning, which is a form of distance learning completely virtualized through digital channels (mainly the internet); and Blended learning, in which e-learning is combined with in-person classes (Romero & Ventura, 2013). The different educational levels are: Primary education, the first stage in formal compulsory education; Secondary education, the final stage of basic education and the phase before tertiary level; and Tertiary education, which refers to education provided mainly at universities, for example leading to academic or professional degrees.

To answer this question, we classified the selected papers by the type of educational system and education level. As Figure 3 shows, the studies used data mostly from online learning (47 papers – 57.3%) followed by traditional in-person environments (30 papers – 36.6%), while very few studies were conducted in hybrid or B-learning environments (5 papers – 6.1%). Figure 3 also shows that most of the 82 papers described studies done with students in tertiary education (76 papers – 86.6%), a few with secondary level students (6 papers – 7.3%), and none with primary level students. This indicates that most of the effort to date has been in early prediction with university students, which is also in accordance with the accessibility of the data. Student data from learning environments is easier to collect, manage and analyse, and in the authors' experience, higher education is much more computerized than primary and secondary education.

Table 2 shows a summary of the 82 selected papers grouped by type of educational environment and education level.

Table 1
Top 5 most cited papers in Google Scholar

#	Title	Reference	#Cites
1	Mining LMS data to develop an "early warning system" for educators: A proof of concept	(Macfadyen & Dawson, 2010)	1028
2	Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative	(Jayaprakash et al. 2014)	332
3	Early Prediction of Student Success: Mining Students Enrolment Data	(Kovačić, 2010)	262
4	Identifying significant indicators using LMS data to predict course achievement in online learning	(You, 2016)	245
5	Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses	(Costa et al., 2017)	199

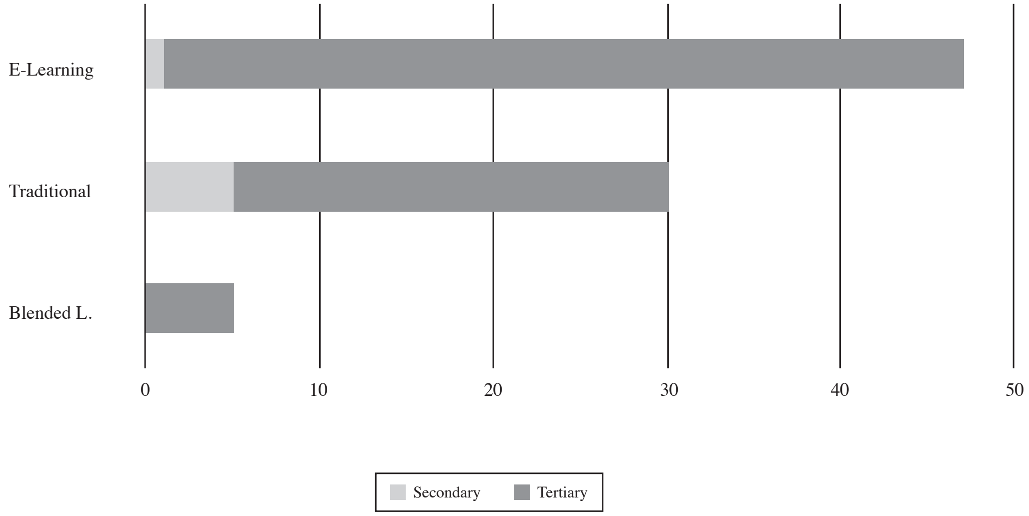

Figure 3. Education level data by type of learning environment

Table 2
Summary of all selected papers by type of educational environment and education level

Educational Environment	Education Level	# Papers	%
Face-to-face	Secondary	5	6.1
	Tertiary	25	30.5
E-Learning	Secondary	1	1.2
	Tertiary	46	56.1
B-Learning	Tertiary	5	6.1

What EDM techniques have been most used to date?

There are different data mining techniques for early prediction of student performance, both supervised (classification and regression) and unsupervised (clustering and association). Classification tries to predict a categorical or nominal value whereas regression tries to predict a numerical value. Clustering puts similar objects into groups and association finds associations or relationships.

Figure 4 shows the frequency of use of techniques in the 82 selected papers in order to determine the most widely-used techniques in EDM. Classification is the most commonly-used technique with 50 papers (42.4%), followed by regression with 33 papers (28%). Clustering, with 13 papers (11%), and association, with 2 papers (1.7%), were used much less often, along with other techniques that were not specified (16.9% noted Machine Learning / Data Mining generically). Hence, the two main DM techniques that have traditionally been applied to early prediction of student academic performance are classification and regression, both supervised techniques. Regression techniques have been used to predict the specific numerical value of a student's performance, and classification has been used to predict the class to which the student belongs, such as Pass/Fail, Success/Failure, or Retain/Dropout.

Which specific algorithms are the most used, and which have produced the best prediction results?

There is a wide range of specific data mining algorithms for doing early prediction. In classification, the most popular were Decision Tree, Random Forest, Support Vector Machine, Naive Bayes, K-Nearest-Neighbour, Boosted Tress, Adaptive Boosting, Gradient Boosting. Popular regression algorithms included Logistic Regression, Linear Regression, and Bayesian Additive Regressive Trees. In Clustering, the popular algorithms were K-Means, Balanced Iterative Reducing, and Clustering using Hierarchies, while in Association, they were Class Association Rule and Random Guess.

Table 3 shows a summary giving the type of DM method, the name of the specific algorithm, and the number of times each algorithm was used in the papers in absolute and percentage terms. The most widely-used algorithms were Naive Bayes, Decision Tree, Support Vector Machine and Logistic Regression.

In terms of algorithm accuracy, the best results were obtained by Miguéis et al. (2018), who achieved 96.1% prediction accuracy with Random Forest, and Razak et al. (2018), who achieved 96.2% with linear regression and 82% with decision tree (J48). Jiang et al. (2014) achieved 92.6% accuracy with logistic regression. Costa

Table 3
Most used algorithms and best results if authors provide them

Method	Algorithm	#	%
Classification	Decision Tree (J48)	31	38%
	Random Forest	25	30%
	Support Vector Machine	21	26%
	Naive Bayes	14	17%
	K-Nearest-Neighbor	10	12%
	Boosted Trees	7	9%
	Adaptive Boosting	7	9%
	Gradient Boosting (XGBoost)	3	4%
	Other	5	6%
Regression	Logistic Regression	23	28%
	Linear Regression	12	15%
	Bayesian Additive Regressive Trees	1	1%
	Other	12	15%
Clustering	K-Means clustering	2	2%
	Balanced Iterative Reducing and Clustering using Hierarchies	1	1%
Association	Class Association Rule	1	1%
	Random Guess	1	1%

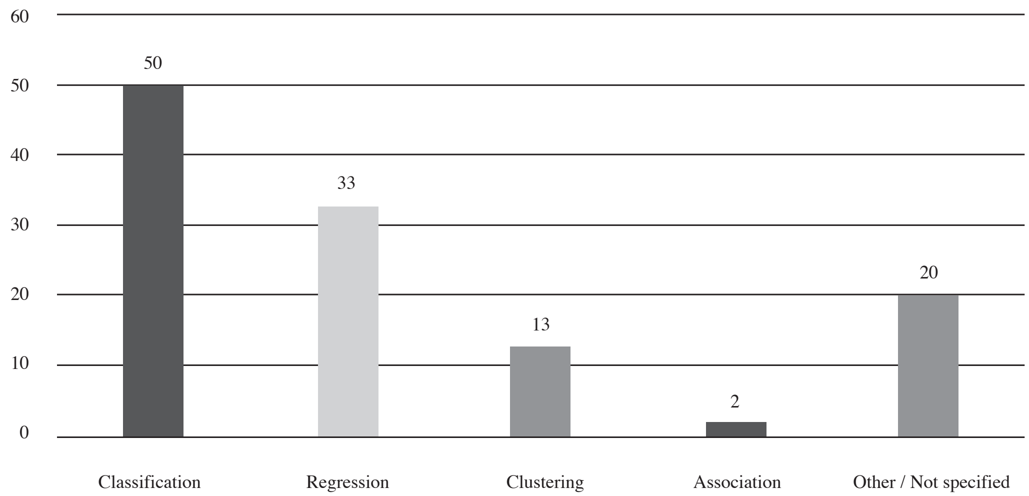

Figure 4. Frequency of use of EDM techniques

et al. (2017) achieved 92% accuracy. However, they also applied naive Bayes and decision tree algorithms as did Casey (2017), who achieved 69% prediction accuracy. In contrast, Chung & Lee (2018) achieved 95% accuracy with their best model applying random forest, while Wang et al. (2018) applied naive Bayes and achieved 85% accuracy.

How much earlier can academic performance be accurately predicted through EDM techniques?

Course length varies depending on the educational environment. For example, traditional education courses can last from four months to a semester or a year. The B-learning modality is similar because the system generally fits the times of an in person (traditional) course, while in e-learning, a course can last from a few weeks to several months. This means that there are different timespans for early prediction, therefore, the answer to this question is addressed for each type of educational environment separately. Early prediction times will depend on the modality of the course.

Traditional Environment

Within the traditional in-person educational environment, most papers do not explicitly indicate how early they can predict academic performance, very few provide that information. Berens et al. (2018) conducted a study over several semesters of bachelor's degrees at two universities (state and private). They showed that the prediction accuracy significantly improved as the semesters went by. At the time of the students' enrolment, they achieved 68% prediction accuracy for the public university and 67% for the private. After obtaining student performance data at the end of the first semester, they achieved 79% accuracy for the public university and 85% for the private, and after the fourth semester, the prediction accuracy reached 90% for the public and 95% for the private. In contrast, Wang et al. (2018) only indicated that success or failure can be predicted in the first semester with good accuracy. Bursać et al. (2019) used models that were, in the second week of a 13-week course, able to determine whether some of the students needed assistance in learning and assimilating learning materials in order to achieve a good grade at the end of the educational process.

E-Learning Environment

One of the most notable of the papers about e-learning courses was from Kuzilek et al. (2015). They managed to increase prediction accuracy by approximately 50% at the beginning of the semester and more than 90% at the end of a high school course. In a 16-week course, Han et al. (2016) produced a model in which the area under the curve, AUC (an indicator of the goodness of the prediction that represents the relationship between the sensitivity and specificity of a predictive model), was in the 0.62-0.83 range, predicting a week ahead. Howard et al. (2018) predicted students' final grades at week 6 (out of 12), based on a mean absolute error up to 6.5 percentage points. Vitiello et al. (2018) achieved 0.8 Accuracy when considering the active time of 10% of the users or the first five days after the initial user interaction. According to Hlosta et al. (2017), it is important for evaluations to be performed in the first few days of a course. If the score is over 50%, there is a high probability of students' academic success. Aljohani et al. (2019) Predicted pass/fail classes with around 90% accuracy within the first 10 weeks of student interaction in a virtual learning environment. Queiroga et al. (2020) predicted at-risk students with an AUC above 0.75 in the initial weeks of a course. Li et al. (2020), reported an AUC score of 0.8262 in the task of next-day prediction while the performance fell to 0.7430 in a next-two-week prediction task.

B-Learning Environment

In papers about B-learning, Costa et al. (2017) achieved an accuracy that varied between 0.50 and 0.82 in a distance education course and from 0.50 to 0.79 for a course on the learning environment. These results indicate that after the first week of these courses, it was possible to identify students who were likely to fail with at least 50% effectiveness. Lu et al. (2018) showed that the final academic performance of students in a blended course could be predicted with high stability and accuracy between weeks 1-6 of the course (out of 18). Macarini et al. (2019) detected at-risk students in the first week of a course with an AUC value from 0.7 to 0.9.

What specific variables or attributes have been used and produced better performance?

The variables and student attributes used for prediction vary depending on the educational environment, and even within the

Table 4

Most used variables classified by educational environment and source of data

FACE-TO-FACE	E-LEARNING	B-LEARNING
Demographics: age, nationality, sex, city, family income level, having a scholarship, having a job or baby, living with parents, legal guardians' educational attainment Activity: Homework grade, homework clicks, attendance, discussion, positive valence, negative valence, neutral valence, average of valence, ePortfolio engagement features Performance: Total credits, credits gained, failing credits, passing rate, arithmetic mean score, weighted average credit score, average credit score point, credit score point, failing score	Interaction: Videos watched, problems attempted; total number of activities; total number of active days; total number of sessions, number of successful compilations, ratio between on-campus and off-campus connections, number of connections, time spent on the platform, time spent on slides within the platform, time spent typing in the platform, time idle in the platform, slides covered, number of slides visited, number of slides opened, number of transactions, number of mail messages read, number of mail messages sent, number of discussion messages read, number of files viewed, number of web links viewed, number of clicks. Performance: number of assessments started, number of assessments finished, time spent on assessments, number of assignments read, number of assignments submitted, time spent on assignments	On-campus: age, gender, civil status, income, number of homework exercises, participation in class, performance in weekly activities and final exam Distance education: time and number of accesses and messages in communication tools (blog, glossary, wiki, and forums), video-viewing behaviour, out-of-class practice behaviour, number of clicks and time with other course resources, quiz scores and virtual tutoring

same environment, the variables vary between studies. Researchers have used different groups of variables in each paper, which makes it hard to tabulate the variables by frequency of use. In general, these variables come from the same data sources, such as student demographics, student activities and student interactions. Table 4 shows the most commonly-used variables in the selected papers grouped by the type of educational system and source of data.

As Table 4 shows, in Traditional education, there are three main sources of variables: demographics, performance, and activity. In E-learning environments there are only two: variables related to student interactions and performance. Finally, on-campus and distance education related variables were found to be used in B-learning systems. In order to see which variables produced the most accurate predictions, we examine each type of educational environment separately below.

Traditional Environments

In traditional in-person educational environments, there are a group of variables that were used most. Berens et al. (2018), Cano & Leonard (2019), and Araújo et al. (2019) used academic performance data and student demographic data to achieve a 79% prediction accuracy at the end of the first semester for a public university and 85% for a private university in applied sciences. Along similar lines, Aguiar et al. (2014) used similar data, supplemented with ePortfolio engagement features, where the highest AUROC value (0.929) was obtained by the dataset with the highest academic participation, and the academic performance was worst with an AUROC value of 0.654. Kovačić (2010) used student demographic data and the study environment to achieve a general classification percentage of 60.5%. Yu et al. (2018) considered the relative variables of tasks, assistance, and discussion. They also considered a variable called courage, which is obtained by applying sentiment analysis to identify affective information within self-evaluations based on written text, comments that reflect learning attitudes towards the lesson, comprehension of the course content, and learning difficulties, which produced a prediction accuracy of 76%.

E-Learning Environments

In e-learning education systems, most of the studies used attributes related to interaction with the learning environment. Kuzilek et al. (2015) used these types of attributes to achieve 93.4% accuracy. Similarly, Chui et al. (2018) used these same types of attributes, among others related to module presentation, and achieved between 92.2% and 93.8% accuracy predicting at-risk students. Among the papers that focused more on the attributes of interaction with the study courses, Han et al. (2016) used attributes such as time of interaction with resources, the interaction of students with problems and submissions, and study habits to achieve an AUC between 0.62 and 0.83. Other studies used attributes such as the number of emails sent, and the number of evaluations made. Macfadyen & Dawson (2010) and Nistor & Neubauer (2010) achieved significant prediction results and they indicated that quiz marks were a very important predictive factor. Olivé et al. (2019) used neural networks to predict which students were likely to submit their assignments on time using data from student and peer activity, student activity and peer activity separated from course info, and student activity, peer activity, and course information trained separately (the networks with the greatest predictive power). Mbouzao et al. (2020) identified failure patterns of up to 60% of students who would dropout or fail the course based on the first week student interaction with MOOC videos in a thirteen-week course, and were able to identify 78% of successful students. Kuzilek et al. (2015), Ortigosa et al. (2019), Kostopoulos et al. (2019), and Waheed et al. (2020) used demographic and variable data from the LMS. Choi et al. (2018), Aljohani et al. (2019), Villa-Torrano et al. (2020), Chen & Cui (2020), and Cui et al. (2020) used the number of clicks as a predictive attribute.

B-Learning Environments

The most used variables for B-Learning environments came from on-campus traditional in-person and distance or e-learning sources. Costa et al. (2017) used attributes such as gender, marital status, age, exam, forums, access, messages, wiki, and transfers, producing predictions that were 92% accurate. Lu et al. (2018) used attributes such as video visualization, out-of-class practice behaviour, homework and questionnaire marks, and after-school tutoring assistance, achieving accuracy between 82-83%. Macarini et al. (2019) used data linked to three different aspects of student interactions (cognitive presence, teaching presence, and social presence) aiming to predict students at risk of failing based on an existing theory about how interactions work inside Virtual Learning Environments. Gitinabard et al. (2019) found that the most important features were total time spent in both types of sessions, total number of actions performed in both browser and study sessions, number of study and browser sessions, number of homogeneous sessions between study and browser sessions.

Research Directions

In this paper, we have described the current state of the art in early prediction of student performance through data mining techniques by means of a systematic review of the literature. We also defined five research questions whose answers can provide important findings for the scientific educational community:

- With regard to the first research question, we have shown that most of the published papers were about online learning systems and traditional in-person secondary and tertiary education. However, very little research has been conducted on early prediction in primary education, which is an open research area. According to the results published in some recent papers, one very promising field is the application of data mining techniques for early prediction of student performance in blended learning environments.
- In relation to the second question, we have shown that the most commonly-used techniques were classification and regression. However, it should be noted that the application of association and clustering in conjunction with the first two may imply a certain trend. At the very least, the clustering technique was shown to be able to be used to make a prediction without using any other techniques (Chau et al., 2018).
- In terms of the third question, we have shown that within each technique, there were some specific algorithms that were widely used and which have produced very good

prediction results. In the classification technique, the stand outs were J48, Random Forest, SVM, and Naive Bayes stand out, while in the regression technique, logistic regression and linear regression stood out. These algorithms are recommended for new researchers when dealing with an early prediction problem.
- With regard to the fourth question, we have shown that how early the prediction can be done varies based on the type of educational system. Within traditional in-person education, Berens et al., (2018) achieved an accuracy of between 78%-84% predicting dropout, with data from the first semester by using average grade (avg. Grade/semester) as the most important predictor. In e-learning environments, an evaluation test should be performed in the first few days of the course, such that if the test score is over 50%, there will be a high probability of a student's academic success (Hlosta et al., 2017).
- In relation to the fifth question, we have shown that most studies used student assessment data when doing early prediction. Within traditional environments, most of the papers also used demographic data to make predictions (Aguiar et al., 2014). Meanwhile, in virtual environments (e-learning and B-learning), most of the variables were gathered from students' interaction with the system and there is an increasing interest in sentiment analysis data (Yu et al., 2018).

Finally, we would like to highlight some future lines that we consider important research opportunities for the EDM research community:

- Selecting and evaluating what are the most important very early factors or indicators that affect student performance in each type of educational system and at each level: More research is needed on selecting the best features to use according to the type of educational system in order to be able to provide earlier predictions (for example in the first day or week, or even before starting the course, when the student registers). This can be dealt with as a multi-view problem, in which the huge amounts of data used for making predictions come from multiple sources and different data sources and we need to select the best attributes.
- Generalizing early prediction models in order to apply them or transfer them to other courses. There is a need to generalize and reuse these models but providing good accuracy is a challenge because they are specific to the courses. The problem is that each study uses different features according to the characteristics of each course, which creates difficulties in adapting any one of the existing plethora of models to any course. More work is necessary to produce good models that are transferable to different courses from the original.
- Developing and testing Early Warning Systems (EWS) and Response to Intervention (RtI) in a real education environment. Real early warning environments should be integrated to close the circle so that following prediction, actions or mitigation measures should be taken for at-risk students at risk: show results, send reports, make recommendations, provide feedback to different stakeholders, etc. More research is necessary in EDM to develoo frameworks, early warning systems and apply real-time intervention strategies in educational environments to work together with educational science (Romero & Ventura, 2019).

Acknowledgments

The authors gratefully acknowledge the financial support provided by the Spanish Ministry of Science and Innovation in the project PID2020-115832GB-I00.

References

Aguiar, E., Ambrose, G. A. A., Chawla, N. V., Goodrich, V., & Brockman, J. (2014). Engagement vs Performance: Using Electronic Portfolios to Predict First Semester Engineering Student Persistence. *Journal of Learning Analytics, 1*(3), 7-33. https://doi.org/10.18608/jla.2014.13.3

Aljohani, N. R., Fayoumi, A., & Hassan, S.-U. (2019). Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability, 11*(24), 7238. https://doi.org/10.3390/su11247238

Ameen, A. O., Alarape, M. A., & Adewole, K. S. (2019). Students' academic performance and dropout prediction. *Malaysian Journal of Computing, 4*(2), 278-303. https://doi.org/10.24191/mjoc.v4i2.6701

Araújo, A., Leite, C., Costa, P., & Costa, M. J. (2019). Early identification of first-year students at risk of dropping out of high-school entry medical school: The usefulness of teachers' ratings of class participation. *Advances in Health Sciences Education, 24*, 251-268.

Barnes, T., Desmarais, M., Romero, C., & Ventura, S. (2009, July). Educational Data Mining 2009 [Conference presentation]. 2nd International Conference On Educational Data Mining, Córdoba, Spain.

Berens, J., Schneider, K., Gortz, S., Oster, S., & Burghoff, J. (2019). Early detection of students at risk-predicting student dropouts using administrative student data from German universities and machine learning methods. *Journal of Educational Data Mining, 11*(3), 1-41. https://doi.org/10.5281/zenodo.3594771

Bogarín, A., Cerezo, R., & Romero, C. (2018). Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs). *Psicothema, 30*(3), 322-329. http://dx.doi.org/10.7334/psicothema2018.116

Bursać, M., Blagojević, M., & Milošević, D. (2019). Early prediction of student success based on data mining and artificial neural network. In D. Milošević, Y. Tang, & Q. Zu (Eds.), Lecture Notes in Computer Science: Vol. 11956. Human Centered Computing (pp. 26-31). Springer. https://doi.org/10.1007/978-3-030-37429-7_3

Cano, A., & Leonard, J. D. (2019). Interpretable Multiview Early Warning System Adapted to Underrepresented Student Populations. *IEEE Transactions on Learning Technologies, 12*(2), 198-211. https://doi.org/10.1109/TLT.2019.2911079

Casey, K. (2017). Using Keystroke Analytics to Improve Pass-Fail Classifiers. *Journal of Learning Analytics, 4*(2), 189-211. https://doi.org/10.18608/jla.2017.42.14

Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying Data Mining Techniques to e-Learning Problems. In L. C. Jain, R. A. Tedman, & D. K. Tedman (Eds.), Studies in Computational Intelligence: Vol. 62. Evolution of Teaching and Learning Paradigms in Intelligent Environment (pp. 183-221). Springer. https://doi.org/10.1007/978-3-540-71974-8_8

Chau, L. M., Chau, V. T. N., & Phung, N. H. (2018). On Temporal Cluster Analysis for Early Identifying In-trouble Students in an Academic

Credit System. In H. T. Bao, L. A. Cuong, H. Van Khuong, & B. T. Lam (Eds.), Proceedings of the 2018 5th NAFOSTED Conference on Information and Computer Science (NICS) (pp. 171-176). IEEE. https://doi.org/10.1109/nics.2018.8606827

Chen, F., & Cui, Y. (2020). Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics, 7*(2), 1-17. https://doi.org/10.18608/JLA.2020.72.1

Choi, S. P. M., Lam, S. S., Li, K. C., & Wong, B. T. M. (2018). Learning analytics at low cost: At-risk student prediction with clicker data and systematic proactive interventions. *Educational Technology and Society, 21*(2), 273-290. https://doi.org/10.2307/26388407

Chui, K. T., Fung, D. C. L., Lytras, M. D., & Lam, T. M. (2018). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior, 107*, 105584. https://doi.org/10.1016/j.chb.2018.06.032

Chung, J. Y., & Lee, S. (2018). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review, 96*, 346-353. https://doi.org/10.1016/j.childyouth.2018.11.030

Cui, Y., Chen, F., & Shiri, A. (2020). Scale up predictive models for early detection of at-risk students: A feasibility study. *Information and Learning Sciences, 121*(3-4), 97-116. https://doi.org/10.1108/ILS-05-2019-0041

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *Artificial Intelligence Magazine, 17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Felix, I., Ambrósio, A. P., LIMA, P. D. S., & Brancher, J. D. (2018, October). Data Mining for Student Outcome Prediction on Moodle: A systematic mapping [Conference presentation]. Anais Do XXIX Simpósio Brasileiro de Informática Na Educação (SBIE 2018), Fortaleza, Brasil. https://doi.org/10.5753/cbie.sbie.2018.1393

Gitinabard, N., Xu, Y., Heckman, S., Barnes, T., & Lynch, C. F. (2019). How Widely Can Prediction Models be Generalized? An Analysis of Performance Prediction in Blended Courses. *ArXiv, 12*(2), 184-197.

Han, W., Jun, D., Xiaopeng, G., Qiaoye, Y., & Kangxu, L. (2017). Predicting Performance in a Small Private Online Course. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), Proceedings of the 10th International Conference on Educational Data Mining (pp. 384-385). Education Resources Information Center.

Hlosta, M., Zdrahal, Z., & Zendulka, J. (2017). Ouroboros: Early identification of at-risk students without models based on legacy data. In A. Wise, P. H. Winne, & G. Lynch (Eds.), Proceedings of the Seventh International Learning Analytics & Knowledge Conference - LAK í17 (pp. 6-15). ACM. https://doi.org/10.1145/3027385.3027449

Howard, E., Meehan, M., & Parnell, A. (2018). Contrasting prediction methods for early warning systems at undergraduate level. *Internet and Higher Education, 37*, 66-75. https://doi.org/10.1016/j.iheduc.2018.02.001

Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics, 1*(1), 6-47.

Jiang, S., Williams, A. E., Schenke, K., Warschauer, M., & Dowd, D. O. (2014). Predicting MOOC Performance with Week 1 Behavior. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), Proceedings of the 7th International Conference on Educational Data Mining (pp. 273-275). IEDMS.

Kostopoulos, G., Karlos, S., & Kotsiantis, S. (2019). Multiview Learning for Early Prognosis of Academic Performance: A Case Study. *IEEE Transactions on Learning Technologies, 12*(2), 212-224. https://doi.org/10.1109/TLT.2019.2911581

Kovačić, Z. (2010). Early Prediction of Student Success: Mining Students Enrolment Data. Proceedings of Informing Science & IT Education Conference (pp. 647-665). https://doi.org/10.28945/1281

Krotseng, M. V. (1992). Predicting persistence from the student adaptation to college questionnaire: Early warning or siren song? *Research in Higher Education, 33*, 99-111. https://doi.org/10.1007/BF00991974

Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrahal, Z., & Wolff, A. (2015). OU Analyse: Analysing at-risk students at The Open University. Learning Analytics Review, LAK15-1, 1-16.

Li, H., Ding, W., & Liu, Z. (2020). Identifying at-risk k-12 students in multimodal online environments: A machine learning approach. arXiv preprint, 2003.09670.

Liz-Domínguez, M., Caeiro-Rodríguez, M., Llamas-Nistal, M., & Mikic-Fonte, F. (2019, June). Predictors and early warning systems in higher education ó A systematic literature review [Conference presentation]. CEUR Workshop at LASI-SPAIN, Vigo, Spain.

Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Lin, A. J. Q., & Yang, S. J. H. (2018). Applying Learning Analytics for the Early Prediction of Students í Academic Performance in Blended Learning. *Educational Technology and Society, 21*(2), 220-232. https://doi.org/10.2307/26388400

Macarini, L. A. B., Cechinel, C., Machado, M. F. B., Ramos, V. F. C., & Munoz, R. (2019). Predicting students success in blended learning-Evaluating different interactions inside learning management systems. *Applied Sciences, 9*(24), 5523. https://doi.org/10.3390/app9245523

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an ìearly warning systemî for educators: A proof of concept. *Computers and Education, 54*(2), 588-599. https://doi.org/10.1016/j.compedu.2009.09.008

Mbouzao, B., Desmarais, M. C., & Shrier, I. (2020). Early Prediction of Success in MOOC from Video Interaction Features. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), Lecture Notes in Computer Science: Vol. 12164. Artificial Intelligence in Education (pp. 191-196). Springer. https://doi.org/10.1007/978-3-030-52240-7_35.

McMahon, B. M., & Sembiante, S. F. (2020). Re-envisioning the purpose of early warning systems: Shifting the mindset from student identification to meaningful prediction and intervention. *Review of Education, 8,* 266-301. https://doi.org/10.1002/rev3.3183

Miguéis, V. L., Freitas, A., García, P. J. V, & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems, 115*, 36-51. https://doi.org/10.1016/j.dss.2018.09.001

Mújica, D., Pérez Villalobos, A., Bernardo Gutiérrez, M. V., Cervero Fernández-Castañón A. B., & González-Pienda García, J. A. (2019). Affective and cognitive variables involved in structural prediction of university dropout. *Psicothema, 31*(4), 429-436. https://doi.org/31.10.7334/psicothema2019.124

Na, K. S., & Tasir, Z. (2018). Identifying at-risk students in online learning by analysing learning behaviour: A systematic review. In A. Selamat (Ed.), Proceedigns of the 2017 IEEE Conference on Big Data and Analytics (pp. 118-123). IEEE. https://doi.org/10.1109/ICBDAA.2017.8284117

Nik Nurul Hafzan, M. Y., Safaai, D., Asiah, M., Mohd Saberi, M., & Siti Syuhaida, S. (2019). Review on Predictive Modelling Techniques for Identifying Students at Risk in University Environment. In L. M. Hee (Ed.), Proceedings of the 2019 MATEC Web of Conferences 255 (03002). EDP Sciences. https://doi.org/10.1051/matecconf/201925503002

Nistor, N., & Neubauer, K. (2010). From participation to dropout: Quantitative participation patterns in online university courses. *Computers and Education, 55*(2), 663-672. https://doi.org/10.1016/j.compedu.2010.02.026

Olivé, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2019). A Quest for a One-Size-Fits-All Neural Network: Early Prediction of Students at Risk in Online Courses. *IEEE Transactions on Learning Technologies, 12*(2), 171-183. https://doi.org/10.1109/TLT.2019.2911068

Ortigosa, A., Carro, R. M., Bravo-Agapito, J., Lizcano, D., Alcolea, J. J., & Blanco, Ó. (2019). From Lab to Production: Lessons Learnt and Real-Life Challenges of an Early Student-Dropout Prevention System. *IEEE Transactions on Learning Technologies, 12*(2), 264-277. https://doi.org/10.1109/TLT.2019.2911608

Queiroga, E. M., Lopes, J. L., Kappel, K., Aguiar, M., Araújo, R. M., Munoz, R., Villarroel, R., & Cechinel, C. (2020). A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course. *Applied Sciences, 10*(11), 3998. https://doi.org/10.3390/app10113998

Razak, R. A., Omar, M., Ahmad, M., & Mara, P. (2018). A Student Performance Prediction Model Using Data Mining Technique. *International Journal of Engineering & Technology, 7*(2.15), 61-63.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*, 12-27. https://doi.org/10.1002/widm.1075

Romero, C., & Ventura, S. (2019). Guest Editorial: Special Issue on Early Prediction and Supporting of Learning Performance. *IEEE Transactions on Learning Technologies, 12*(2), 145-147. https://doi.org/10.1109/TLT.2019.290810

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355

Costa, E. B., Fonseca, B., Almeida-Santana, M., Ferreira de Araújo, F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of studentsí academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247-256. https://doi.org/10.1016/j.chb.2017.01.047

Siemens, G. (2013). Learning Analytics: The Emergence of a Discipline. *American Behavioral Scientist, 57*(10), 1380-1400. https://doi.org/10.1177/0002764213498851

Tranfield, D., Denyer, D. & Smart, P. (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management, 14*, 207-222. https://doi.org/10.1111/1467-8551.00375

Villa-Torrano, C., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., & Gómez-Sánchez, E. (2020). Early prediction of studentsí efficiency during online assessments using a long-short term memory architecture. In A. Martínez-Monés, A., Álvarez, M. Caeiro-Rodríguez, & Y. Dimitriadis (Eds.), Proceedings of Learning Analytics Summer Institute Spain 2020 (pp. 39-46). CEUR Workshop Proceedings.

Vitiello, M., Walk, S., Helic, D., Chang, V., & Guetl, C. (2018). User behavioral patterns and early dropouts detection: Improved users profiling through analysis of successive offering of MOOC. *Journal of Universal Computer Science, 24*(8), 1131-1150.

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior, 104*, 106-189. https://doi.org/10.1016/j.chb.2019.106189

Wang, Z., Zhu, C., Ying, Z., Zhang, Y., Wang, B., Jin, X., & Yang, H. (2018). Design and Implementation of Early Warning System Based on Educational Big Data. In W. Tu, L. Wang, C. Ji, N. Chen, Q. Sun, X. Song, & X. Wang (Eds.), Proceedigns of the 5th International Conference on Systems and Informatics (pp. 549-553). IEEE. https://doi.org/10.1109/icsai.2018.8599357

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education, 29*, 23-30. https://doi.org/10.1016/j.iheduc.2015.11.003

Yu, L. C., Lee, C. W., Pan, H. I., Chou, C. Y., Chao, P. Y., Chen, Z. H., Tseng, S. F., Chan, C. L., & Lai, K. R. (2018). Improving early prediction of academic failure using sentiment analysis on self-evaluated comments. *Journal of Computer Assisted Learning, 34*, 358-365. https://doi.org/10.1111/jcal.12247