

A Critical Review of Data Mining for Education: What has been done, what has been learnt and what remains to be seen

Ilias PAPADOGIANNIS, Vassilis POULOPOULOS, Manolis WALLACE

ARTICLE INFO

Article History:

Received 18.11.2019

Received in revised form

19.06.2020.

Accepted

Available online 01.10.2020

ABSTRACT

This article provides a thorough review of educational data mining (EDM) in the period 2015-2019. Going beyond earlier review works, in this article we examine previous research from a variety of aspects, including the examined data, the algorithms used, the type of conclusions drawn, the educational level/setting of application and the actual exploitation of the results in the educational setting. Our findings indicate that tertiary education dominates the EDM domain, while minimal focus has been given to secondary education and almost none to primary education. Our finding, and suggestion, is that by focusing EDM on earlier education level the field can have a more profound impact on education and on society as a whole.

© IJERE. All rights reserved

Keywords:

Educational data mining, student performance, secondary education, tertiary education, methods, accuracy

INTRODUCTION

The use of data mining techniques in educational data has increased enormously in recent years. This has been pushed in part from the huge increase in the amount of educational data that is now available. The introduction of information systems enables the recording and retention of large volumes of data in educational institutions. The development of modern as well as asynchronous distance learning has also increased the volume and the variety of types of data. Thus, the conditions for the application of data mining techniques in education have been created and educational data mining has evolved into a distinct interdisciplinary discipline. (Romero & Ventura, 2007), (Papamitsiou & Economides, 2014), (Peña-Ayala, 2014), (Thakar et. al., 2015), (Sukhija et. al., 2015), (Del Río & Pineda Insuasti, 2016).

There have been various attempts to categorize the scientific field. According to Agasisti and Bowers (2017) there are three distinct approaches in data mining in the field of education: a) Educational Data Mining (EDM), b) Learning Analytical (LA) and c) Academic Analytical (Ac An). EDM uses data mining techniques to understand patterns and iterations. LA integrates many of the EDM models while focusing on teaching and learning. In this sense, its main aim is to better inform teaching practices. Ac An focuses on the organizational level and the improvements that can be made to the training process and its results.

The three approaches are difficult to distinguish, as tools, research questions, and administrative use of findings tend to overlap, leading many in the literature to refer to all of them cumulatively as EDM. In any case, this classification can be considered "tentative" or "early" as the scientific field is still rapidly evolving (Agasisti & Bowers, 2017). In this work we extensively examine work published in the broader area of data mining in education over the last few years, focusing particularly in works that aim to assess and predict student performance. We identify trends and we discuss what the future may hold for this exciting field and which promising directions remain to be explored.

Methods

In relation to the applied methods, Baker and Yacef (2009), Romero and Ventura (2010), Agasisti & Bowers (2017) suggest the following methods of EDM.

Prediction: Where the goal is to develop a model that predicts the value of a variable from a combination of other independent variables. Classification techniques or different types of regressions can be used.

Distillation of data for human judgment: This method summarizes and presents information using a useful, interactive, and visually appealing modality, in order to understand a large amount of educational

¹ Corresponding e-mail: i.papadogiannis@uop.gr, orcid.org/0000-0002-7758-0571,

² vacilos@uop.gr, orcid.org/0000-0003-1707-3153

³ wallace@uop.gr, orcid.org/0000-0002-4629-5946

1,2,3, FAB LAB - Knowledge and Uncertainty Research Laboratory, Department of Informatics and Telecommunications, University of the Peloponnese, Greece

data and to better support decision making. It helps users understand information (teachers, education executives).

Relationship mining: The goal is to discover the relationships between variables. This can take the form of trying to find out which forms are interrelated and to what extent. In general, there are four types of relationship mining: association rule mining, correlation mining, sequential pattern mining and causal data mining.

Discovery with models: In this method, a model developed through machine learning, is later used as a tool in another analysis. The model created allows further analysis between the model estimates and the value observations of the variables to be studied.

Outlier Detection: The process of detection is the identification of data points do not fit with other in a data set. This data may concern students with learning difficulties or teachers with a completely different behavior from the usual etc.

Social Network Analysis: which maps the flow of relationships and changes in relationships between the entities involved in knowledge (students, educators).

Text Mining: Considering large collections of written resources for generating new information and converting text into structured data for further analysis.

Prior literature reviews

There have been several literature reviews on EDM in general, focusing primarily on research objectives, methods, and predicting variables used.

Table 1. Previous literature reviews

Reference	Year	Findings
Romero and Ventura	2007	It presents the most relevant work until 2005. It categorizes surveys according to their purpose. Recording an increase in the number of enrollments each year. Possible future trends in the field are outlined. It refers to the need to integrate data mining into the educational environment and its importance for researchers and external users.
Baker and Yacef	2009	It identifies the increasing rates of growth in the field. It detects the increase in available data. It presents the change in the goals and methods followed. It notes that there is considerable room for growth (EDM) as well as the importance of open educational data.
Romero and Ventura	2010	It categorizes potential users of the effects. It categorizes tasks based on the techniques applied and their goals. It considers as important issues for the future the ease of use of non-specialist effects in DM, the need for integration in a simple environment and the need for standardization of data and models.
Papamitsiou and Economides	2014	The papers categorized according to methods, objectives, learning settings. They highlighted the major directions of the field. The added value of EDM/LA discussed. The integration of technologies such as game -based and mobile learning also suggested. A SWOT analysis is conducted.
Pena-Ayala	2014	A detailed presentation of articles by method, algorithms used and objectives was done. Statistical and clustering methods were also applied and two patterns of EDM approaches have been identified. A SWOT analysis was conducted and the great potential of the widespread use of information systems in education has also been presented.
Thakar, Mehta and Manisha	2015	Some key areas were identified, such as identifying weak students and predicting student failure, assessing students in specific courses, assessing students' understanding. The lack and need for a unified approach was highlighted.
Sukhija et al.	2015	It presents tools, the techniques and outcomes of various studies from 2001 to 2015. Some challenges for EDM have been identified, such as the lack of coherent data sets at the level of the overall education system, the lack of flexibility in data sets and the lack of trust of educational authorities in EDM results. It was also emphasized that most of the researches were limited to small-scale experiments.
Del Rio & Insuasti	2016	Focused in predicting students' academic performance in higher education from 2011 to 2016. They presented the main methods, predictors and software used by the authors.

The vast majority of the articles reviewed focus on higher education; primary and secondary education do not seem to have attracted the interest of researchers. In this review we will attempt to confirm this quantitatively, examining articles that aim to predict the academic performance of students in different education levels.

Most reviews dealt with the technical purposes and recorded in detail the methods, techniques and objectives of the research. One aspect that has not yet been assessed quantitatively is that of linking research results to educational policy. Such a connection can range from the micro level (class level) to the macro level (educational system level). This review will examine whether the research results have been used using technical data mining to make decisions by the education administration. A detailed model for linking EDM to decision making at various levels of educational policy is presented by Agasisti & Bowers (2017).

Assessing student performance

Student performance assessment is an area of EDM that has gained the attention of several researchers in recent years. A considerable number of researches have been published, mainly on higher education (Bydžovská, 2016). Undoubtedly, the importance of early diagnosis is of great importance since it could potentially be used to develop actions and policies by educational institutions. In attempting to predict academic performance, a significant number of data mining techniques and explanatory features are used, which relate to students' demographic and academic behavioral characteristics, their online behavior etc (Del Río & Pineda Insuasti, 2016). The review of the research of the last five years is the main intention of this article.

Literature review scope and methods

The purpose of this article is twofold: 1) To identify and present research published over the last five years (2015-2019) in relation to assessing students' academic performance using data mining techniques and 2) to identify cases studies of application of research results to decision-making at the institutional level as well as educational policy in general.

Following the steps of PICO methodology (Pai et al., 2004), borrowed from medical research, the following criteria were used to set up this literature review:

Population: Articles on academic performance using data mining techniques in secondary schools, higher education institutions as well as related on line tools.

Intervention: Methods, algorithms, features.

Comparison: Between different evaluation measures.

Outcome: Most commonly used methods, algorithms, features by authors. Applications of findings in the educational field.

The research questions were:

Q1: What are the methods and algorithms used to evaluate academic performance?

Q2: What is the performance per method?

Q3: What are the features used to evaluate academic performance?

Q4: To what extent were the results of the research used for decision making, according to the authors?

Search strategy

The following databases were searched for articles to examine: Springer Link, IEEE Xplore, ERIC, ACM, Science Direct, Journal of Educational Data Mining (JEDM), and also Google Scholar. Articles were originally searched through the Google Scholar search engine, which provides the ability to locate articles in a large number of scientific journals. Additional search has been performed using the corresponding tools from each database. We have checked the keywords in relevant papers and identified the alternative spellings of our search terms.

The search parameters that were used are:

- Search terms: (educational data mining) AND (tertiary education OR secondary education) AND (student performance OR student dropout).
- Time interval: 2015-2019 (plus early 2020).

Following the methodology of Kitchenhams (2010), we developed the following and exclusion criteria.

Inclusion criteria

- Works aimed at predicting students' academic performance in secondary, tertiary and related online education.
- Works that do not use data mining techniques for the prediction of students dropout in secondary, tertiary and related online education..
- Papers published in scientific journals or conference proceedings.
- Works that provide detailed information on the methodology used.

Exclusion criteria

- Works that do not use data mining techniques for the prediction of students' academic performance or dropout.
- Works that do not provide detailed information about the type of attributes used.
- Works that do not provide detailed information on the algorithms used.
- Works that do not provide detailed information for evaluating of the algorithms used.

The papers identified in the first phase were 564. Subsequently, the title and abstracts of the articles were filtered by applying inclusion criteria. This process rejected 439 articles and 125 articles remained to further study of the full text. After studying all the texts and fully applying the inclusion and exclusion criteria, finally, 120 articles emerged, based on which the full literature review was developed. Table 2 summarizes the sources of the articles we studied and the number of articles from each source.

FINDINGS

Research domains

The vast majority of articles were on higher education. This may be due to better access to data through the development of Learning Management Systems (LMS) in higher education institutions, as well as the fact that scientific experimentation can be performed more easily in higher education. As can be seen below, the percentage of research related to universities or colleges reached 78.69 %. This was followed by research in secondary education at 14.75%, while less research was on online platforms. It was examined whether the methods, algorithms and attributes differ between the educational levels/settings and no differentiation was found. On the contrary, in relation to the data used, we observe that online platforms surveys were dominated by online data.

Table 2. Literature review sources

Source	N
IEEE Xplore	29
Springer	25
ScienceDirect	9
ACM Digital Library	9
ERIC	5
Taylor & Francis	3
JEDM (Journal of EDM)	2
Others	43

Table 3. Educational settings

Educational setting	Count	%
Tertiary education	96	78,69%
Secondary education	18	14,75%
MOOC	8	6,56%
Total	122	100,00%

Algorithms

More than one algorithm was used in most articles. For the purposes of our review, the algorithms used by the researchers were categorized in terms of methodology: Association Rules, Other Decision Rules, Bayesian Methods, Decision Trees, Ensemble Methods, Instance Learning, Logistic Regression, Neural Networks, Support Vector Machine and Other Methods. Regression Techniques also used, the measures used and the average value of them were also recorded. Table 4 summarizes the frequencies of the measurement tools used. The accuracy was the most used evaluation tool.

Table 4. Algorithms' Descriptive Statistics (evaluated with accuracy)

Methods	Count	Accuracy	Articles
Association Rules	5	Various Measures	Bhegade and Shinde (2016), Hussain et al.. (2018), S. Abu-Oda and M. El-Halees (2015) Saravanan and Jyothi (2019), Satyanarayana and Nuckowski (2016)
Decision Rules	23	0.78	Ahadi et al.. (2015), Chango et al.. (2019), Daud et al.. (2017), Livieris et al.. (2016), Livieris et al.. (2019b), Livieris et al.. (2019a), Mhetre and Nagar (2017)
Bayesian	51	0.78	Adekitan and Salau (2019), Ahadi et al.. (2015), Ahmad et al.. (2015), Asif et al.. (2017), Athani et al.. (2017), Bergin et al.. (2015), Bhegade and Shinde (2016), Bydzovska (2016), Canagareddy et al.. (2019), Castro-Wunsch et al.. (2017), Chanlekha and Niramitranon (2018), Costa et al.. (2017), Daud et al.. (2017), Francis and Babu (2019), Govindasamy and Velmurugan (2017), Guo et al.. (2016), Hasheminejad and Sarvmili (2018), Huang et al.. (2019), lam-On and Boongoen (2015), Karaal and Ahuja (2019), Kaur and Singh (2016), Kiu (2018), Livieris et al.. (2016), Livieris et al.. (2019b), Livieris et al.. (2019c) Livieris et al.. (2019a), Mahboob et al.. (2017), Mhetre and Nagar (2017), Miguelis et al.. (2018), Moscoso-Zea et al.. (2019), Namomsa and Sharma (2018), Pandey and Taruna (2016), Santoso and Yulia (2019), Saravanan and Jyothi (2019), Satyanarayana and Nuckowski (2016), Shrivass and Tiwari (2017), Singh and Kaur (2017), Strecht et al.. (2015), Yehuala (2015), Zhou et al.. (2015)
Decision Tree	107	0.82	Abdulazeez and Abdulwahab (2019), Adekitan and Salau (2019), Afeni et al.. (2019) Ahadi et al.. (2015), Ahmad et al.. (2015), Altujjar et al.. (2016), Amaya et al.. (2015), Angiani et al.. (2019), Asif et al.. (2017), Bhegade and Shinde (2016), Buenano-Ferndndez et al.. (2019), Bydzovska (2016), Canagareddy et al.. (2019), Caetro-Wunsch et al.. (2017), CEME (2017), Chango et al.. (2019), Comendador et al.. (2016), Costa et al.. (2017), Daud et al.. (2017), Francis and Babu (2019), Govindasamy and Velmurugan (2017), Hasheminejad and Sarvmili (2018), Hassan et al.. (2019), Huang et al.. (2019), lam-On and Boongoen (2015), Kamal and Ahuja (2019), Kaur and Singh (2016), Ketui et al.. (2019), Kiu (2018), Koprinska et al.. (2015b), Koprinska et al.. (2015a), Kostopoulos et al.. (2015), Kumar et al.. (2019), Livieris et al.. (2016), Livieris et al.. (2018), Livieris et al.. (2019b), Livieris et al.. (2019c) Livieris et al.. (2019a), Mahboob et al.. (2017), Mhetre and Nagar (2017), Miguelis et al.. (2018), Moscoso-Zea et al.. (2019), Namomsa and Sharma (2018), Nunez et al.. (2018), Pandey and Taruna (2016), Parmar et al.. (2015), Polyzou and Karypis (2018), Rodrigues et al.. (2019), Ruby and David (2015), Santoso and Yulia (2019), Sara et al.. (2015), Saravanan and Jyothi (2019), Satyanarayana and Nuckowski (2016), Shrivass and Tiwari (2017), Singh and Kaur (2018), Sivakumar and Selvaraj (2018), Strecht et al.. (2015), Tegegne and Alemu (2018), Yehuala (2015)
Ensemble Learning	39	0.81	Abdulazeez and Abdulwahab (2019), Adekitan and Salau (2019), 7, Angiani et al.. (2019), Castro-Wunsch et al.. (2017), CEME (2017), Chanlekha and Niramitranon (2018), GonzalezMarcos et al.. (2019), Hassan et al.. (2019), Huang et al.. (2019), Hussain et al.. (2018b), Hussain et al.. (2018a), Hussain et al.. (2019) Ketui et al.. (2019), Kiu (2018), Knowles (2015), Kostopoulos et al.. (2015), Livieris et al.. (2016), Livieris et al.. (2018), Livieris et al.. (2019c), Mahboob et al.. (2017), Migueis et al.. (2018), Moscoso-Zea et al.. (2019), Pandey and Taruna (2016), Polyzou and Karypis (2018), Satyanarayana and Nuckowski (2016), Shrivass and Tiwari (2017), Zhang et al.. (2018)
Instance Based	19	0.70	Abdulazeez and Abdulwahab (2019), Al-Shehri et al.. (2017), Asif et al.. (2017), Bergin et al.. (2015), Brinton and Chiang (2015), Bydzovskd (2016), Castro-Wunsch et al.. (2017), Chango et al.. (2019), Daud et al.. (2017), Govindasamy and

			Velmurugan (2017), Hasheminejad and Sarvmili (2018), lam-On and Boongoen (2015), Livieris et al.. (2016), Livieris et al.. (2018), Livieris et al.. (2019c), Strecht et al.. (2015)
Logistic Regression	11	0.81	Adekitan and Salau (2019), 7, Bergin et al.. (2015), Burgos et al.. (2018), Canagareddy et al.. (2019), Daud et al.. (2017), He et al.. (2015), Klusener and Fortenbacher (2015), Rodrigues et al.. (2019), Strecht et al.. (2015), Zhang et al.. (2018)
Neural Networks	33	0.79	Adekitan and Salau (2019), 7, Altai et al.. (2019), Amirhajlou et al.. (2019), Asif et al.. (2017), Bergin et al.. (2015), Burgos et al.. (2018), Chanlekha and Niramitrannon (2018), Costa et al..(2017), Francis and Babu (2019), Guo et al.. (2016), Hasheminejad and Sarvmili (2018), Hassan et al.. (2019), Huang et al.. (2019), Hussain et al.. (2018a), Hussain et al.. (2019) lam-On and Boongoen (2015), Kiu (2018), Kumar et al.. (2019), Livieris et al.. (2016), Livieris et al.. (2018), Livieris et al.. (2019c) Livieris et al.. (2019a), Luo et al.. (2015), Naser et al.. (2015), Nunez et al..(2017), Okubo et al.. (2017), Rivas et al.. (2019), Ruby and David (2015), Shrivass and Tiwari (2017), Strecht et al.. (2015), Zhang et al.. (2018)
Support Vector Machines	29	0.78	Al-Shehri et al.. (2017), Amirhajlou et al.. (2019), Angiani et al.. (2019), Bergin et al.. (2015), Bydzovska (2016), Chanlekha and Niramitrannon (2018), Corrigan et al.. (2015), Costa et al..(2017), Daud et al.. (2017), Francis and Babu (2019), Gonzalez-Marcos et al.. (2019), Guo et al.. (2016), Hasheminejad and Sarvmili (2018), Hassan et al.. (2019), Huang et al.. (2019), Kamal and Ahuja (2019), Kumar et al.. (2019), Livieris et al.. (2016), Livieris et al., (2018), Livieris et al.. (2019b), Livieris et al.. (2019c) Livieris et al.. (2019a), MiguSis et al.. (2018), Nunez et al..(2017), Polyzou and Karypis (2018), Rodrigues et al.. (2019), Sara et al.. (2015), Shrivass and Tiwari (2017), Son and FHijita (2019), Strecht et al.. (2015), Zhang et al.. (2018)
Proposed and other algorithms	11	0.77	Bhegade and Shinde (2016), Burgos et al.. (2018), Bydzovska (2015), Francis and Babu (2019), Govindasamy and Velmurugan (2017), Hussain et al., (2018a), Hussain et al.. (2018b), Hussain et al.. (2019) Mai et al.. (2019), Saravanan and Jyothi (2019), Singh and Kaur (2018), Son and Fujita (2019)
Linear Regression	11	Various Measures	Adekitan and Salau (2019), Almasri et al.. (2019), Amirhajlou et al.. (2019), Bydgvoskd (2016), Comendador et al.. (2016), Mai et al., (2019), Nunez et al.. (2018), Okubo et al., (2017), Sorour et al.. (2015), Strecht et al.. (2015), Ulloa-Cazarez et al., (2018)

We have found a high frequency of Decision Tree algorithms. To this high-frequency contribute the widespread use of the method in general and the large number of algorithms available in popular tools, such as WEKA. Bayesian algorithms, mainly Naive Bayes, were also widely used. More advanced methods such as Neural Networks, Support Vector Machines and Ensemble Learning Methods have been used to a lesser extent. Logistic regression, although was quite accurate, was applied in few cases, and the same goes to Instance-Based Learning as well. Next, we present in more detail the descriptive statistics of the most frequently applied algorithms. Due to the use of different evaluation tools, the statistics of those who used the accuracy are presented.

More in detail, Bayesian algorithms were used in a large percentage of the researches we studied. Naive Bayes algorithm was used more often (45 times), and in a few cases Bayesian Networks were used. Frequently, the Naive Bayes algorithm has been used as a benchmark to compare accuracy with other, more complex algorithms. The satisfactory accuracy of the algorithm was confirmed, despite the speed of calculation and low consumption of resources. In five cases it was found that Naive Bayes had the highest accuracy (Bergin et al., 2015; Zhou et al. 2015; Kamal and Ahuja 2019; Kumar et al. 2019; Santoso and Yulia 2019), while the average accuracy of the algorithm reached 0.7560. It should be noted that the required hypothesis of independence between variables cannot be safely assumed or easily manufactured in educational data.

Decision Tree algorithms made up the majority of the algorithms used with the highest accuracy. Of course, there are many different algorithms in the category, such as: C4.5, ID3, CART, Random Trees, etc., with the result that researchers can choose for many different approaches, in many cases in the same article.

The C4.5 algorithm typically showed the highest accuracy. The ADTree and RepTree algorithms were also highly accurate, although they were not used frequently.

A similar score in accuracy compared to Decision Trees was shown by the logistic regression, which is a powerful algorithm, although it was used in only a small number of articles. Instance based learning algorithms, SVM and other Neural Networks showed lower average accuracy, although the comparison between different methods in different fields is relatively problematic (a case of comparison of apples and oranges). In any case, they were less used by researchers.

Table 5. Anova test: Accuracy between Methods

	Sum of Squares	df	Mean Square	F	P
Between Group	0,567	29	0.20	0.786	0.778
Within Group	6,245	251	0.25		
Total	6,812	280			

Ensemble Methods techniques were used in a fairly large number of articles and the Random Forest algorithm was used with an accuracy of 0.79. Stacking and AdaBoost algorithms presented better accuracy, although they were not used as often. We also noticed the very low frequency of using Unsupervised learning techniques such as K-means that was only used in two cases.

A small number of studies used various types of linear regression. Of the eleven cases examined in this review, three were the Coefficient of Determination (R²) and the F-measure was used to evaluate the fitting in data. The mean squared error (MSE) was used four times.

By maintaining limitations about the ability to compare the accuracy of different algorithms in different domains, we checked if an algorithm had a higher accuracy score. Given the normality of the accuracy score distribution, we used the One Way Anova test. As we can see in Table 5, there is no statistically significant difference between the performances of the different categories of algorithms [F=1.143, p=0.334].

In Table 6, we can see the frequencies of using each evaluation tool. The vast majority of authors preferred accuracy score.

Attributes used

The results of the algorithms were tested with several different measures. Table 6 shows the frequencies of three most used attributes from the articles have used accuracy as an evaluation measure (grades, demographics and academic data besides grades).

The largest percentage of surveys we studied used student grades as explanatory factors (33.94%). This significant proportion of research utilized student grades in previous courses or in the early stages of semesters to assess academic performance.

Student demographics in order to assess students' final academic performance. Such were usually the gender, the profession of the parents, the age, the area, etc. In combination with the demographic data, socio-economic data were used, which concerned the financial background of the students and the families from which they come (García-Vélez et al., 2016).

Other academic data of students were also used, in addition to grades, in a significant percentage (29.50%). Student activity data dominated in the research on student performance on online learning platforms. The platforms involved MOOCs as EDx and distance learning platforms from universities e.g. Amrieh et al., (2015), Tang et al., (2015), Mahboob et al., (2017), Hassan et al., (2019). In addition, the scores in various tests were used as well as the demographic data of the students.

Finally, other factors that were used to a much lesser extent were behavioral data, internet records (logs), motivational data, while in two surveys the use of document data was used Zhou et al., (2015).

Table 6: Measurement tool

Measure	Frequency
Accuracy	79,48%
F1 measure	12,50%
Error Rate	1,18%
AUC	1,18%
Precision	0,24%
R ²	0,71%
Mean Squared Error (MSE)	2,59%
Root Mean Square Error (RMSE)	2,12%

We also tested the variance of average accuracy in terms of using demographic socioeconomic, academic data and grades as features. We used these categories of features because they had the highest frequency of use in the articles we reviewed. In Table 7 we observe the high accuracy offered by the use of the score as a predictive factor (from 0.79 to 0.86). In contrast, not using grades, but only demographic and other academic data reduces accuracy (0.77). Using demographic data alone seems to be the most accurate, but it only applies to three cases. Respectively, in five surveys, only academic information was used, with an accuracy of 0.88. Of the studies included in this analysis, which used accuracy as an algorithm evaluation metro, 12 did not use any of the above features. As we can see from the use of the confidence interval, none of the feature combinations show a statistically significant difference in accuracy.

Table 7. Descriptive statistics per attribute

Grades	Academics	Demographics	N	Mean	Lower	Upper	Articles
NO	NO	NO	12	0,82	0,74077	0,89923	<i>a</i>
		YES	3	0,94567	0,81648	1,07485	<i>b</i>
	YES	NO	5	0,8828	0,7848	0,9808	<i>c</i>
		YES	20	0,76905	0,70851	0,82959	<i>d</i>
YES	NO	NO	25	0,79476	0,74941	0,84011	<i>e</i>
		YES	13	0,86154	0,79279	0,93028	<i>f</i>
	YES	NO	8	0,80963	0,73669	0,88256	<i>g</i>
		YES	8	0,80575	0,66823	0,94327	<i>h</i>

a. Articles: Tang et al.: (2015), Klusener and Fortenbacher (2015), Brinton and Chiang (2015), Shrivastava and Tiwari

b. Articles: Al Shehri et al.: (2017), Amaya et al.: (2015), Widyaningsih et al.: (2019)

c. Articles: Okubo et al.: (2017), Singh and Kaur (2018), Santoso and Yulia (2019), Sumitha et al.: (2016), Al Barrak and Al Razgan

d. Articles: Amrieh et al.: (2015), Ruby and David (2015), Almasri et al.: (2019), Sivakumar et al.: (2016), Kumar et al.: (2019), Chanlekha and Niramitranon (2018)

e. Articles: Livieris et al.: (2018), Angiani et al.: (2019), Jishan et al.: (2015), Livieris et al.: (2019c), Al Saleem et al.: (2015), Livieris et al., (2019)

f. Articles: Kostopoulos et al.: (2015), Athani et al.: (2017), Sara et al.: (2015), Kasthuriarachchi and Liyanage (2019), Namomsa and Sharma (2018)

g. Articles: Bergin et al.: (2015), Navamani and Kannammal (2015), Ketui et al. (2019), Bhegade and Shinde (2016), Pristyanto et al., (2018)

h. Articles: Lopez Guarin et al.: (2015), Pradeep et al.: (2015), Yehuala (2015), Kaur and Singh (2016), Mahboob et al.: (2017), Pereira et al., (2018)

Using the findings of data mining

The results of the data mining process may be used for academic/research purposes but may also find direct application in the actual educational practice. Romero and Ventura (2007) reported the following potential uses of EDM research results.

- Students often need alternative learning options and learning activities to improve performance. Feedback and suggestions are also provided so that students can adjust their study or choose new courses or lessons.
- Teachers and trainers can be helped by EDM to develop differentiated teaching methods through student categorization as well as new teaching methods. It also helps in understanding the cognitive, behavioral and social aspects of learning.
- Researchers aim to evaluate methods and algorithms in solving specific educational problems. They also seek to develop new techniques and tools.
- Administrators seek to make better use of the institution's resources, evaluate curricula, and implement new forms of teaching. For all of the these objectives, EDM can help by providing evidence to support decisions.

The using of research results for decision making can be very helpful for education authorities. Evidence-Based decision making enables strong decision support at local, regional, national and even supranational level. We studied the cases in which the results of the studies were used in decision making, according to what is mentioned in the article. From the study of the articles we found that for the most part these are case studies in a limited sample. The main interest of the researchers was to evaluate the effectiveness of specific algorithms and techniques and not to use their results to conduct educational policy. However, there have been a few cases in which a tool has been developed for use by schools or universities. We found that in one case the results, the proposed algorithms but also the whole methodology of the article are still in use for decision making. In five articles, supporting tools for educators are developed. We also tried to contact the authors to get more information about the using of their findings. The feedback received is shown below.

The first article we refer to is by Knowles (2015) and it is probably the most comprehensive one. It concerns the state of Wisconsin which has one of the highest graduation rates in the USA, but large disparities among students. The aim was to create an early warning system (DEWS) for the risk of student dropout. The system covers 225,000 students in Wisconsin. It uses many different algorithms examines demographic, socio-economic and academic data. It trains several algorithms and selects the most efficient ones using "caretEnsemble" package. Ensemble Models showed an AUC of 0.86 to 0.91. The article is a presentation of a complete tool in use. The developed tool and methodology is still in use in Wisconsin State.

Also, in two articles the findings were used to reduce the failure rate of students; specific and measurable results were achieved. In Corrigan et al. (2015) the researchers tried to predict students' performance at the end of the semester. They used data from the Dublin City University Virtual Learning Environment (Moodle). Data such as the number of logs per student, the average residence time, the use on weekends, etc. were used, and a weekly communication and feedback process was established with the students (1,200 students). The SVM algorithm was used. The study's results were used as feedback of the students to help them increase their efforts led to an overall increase of 2.67% in the final grade.

In Burgos et al. (2018), logistic regression was used in historical student performance data to predict students' success or failure. The experiment was conducted on data from 100 students of the Madrid Open University. The data were related to scores on various tests on a distance learning platform. The result showed that the logistic regression technique achieves accuracy above 90% after four weeks data. Anova test was also performed on the model accuracy results for previous years, in order to study the possible changes that have occurred due to the different way in which students are accepted. Although this research is a single case study it is an extremely important idea to apply it in different time frames. The use of the results of the research in decision making in the next academic year, contributed to the reduction of the dropout by 14%.

A java tool for using by the school unit or university institute was created based on the findings of the next two articles. In Livieris et al., (2016) presented a decision support tool for predicting student performance in the final exams of the school year. The data refer to performance in the Mathematics in high school students of 14-15 years old and concern 279 students from a private school in Greece. The algorithms used were Naive Bayes, BP, RIPPER, KNN, C4.5, SMO. The accuracy of these algorithms ranged between 51.6% and 70.3%. After using ensemble techniques, the accuracy reached 90.3% (voting). As the authors

state, this is a pilot study using a small sample. But they proposed a Java based tool for prediction of student's performance.

A similar java tool was developed in Livieris et a (2019). In this interesting research data related to secondary education are used. The researchers were investigated the possibility of identifying students who failed in the courses but also the evaluation of the final grades, in Secondary education. There were grades data for 2260 students in courses of the first 2 years of senior high school. Many algorithms were used, namely MLP, RBF, SMO, NB, C4.5, JRip, KNN (k = 3, 5, 10), PART and Voting. The SMO algorithm was more accurate 91.14%. Finally, final grade forecasting Java based was developed based on the results in order to take appropriate action to improve performance. Unfortunately, we don't have quantitative data about the results of these studies in improvement of the students' performance

Finally, in Fernandes et al. (2019) students' academic performance in Brazilian public schools is envisaged. Two sets of data were used, including demographic and academic data. The first contained variables taken before the start of the course and the second contained variables collected two months after the start. The key demographic attributes were collected in total 238,575 records in the year 2015 and 247,297 records in the year 2016. The algorithm used was the Gradient Boosting Machine. The most important attributes were identified and very high accuracy was calculated. The researchers mention that they aim to use their methodology by the regional education authorities, but that is not yet possible, as the first author informed us.

DISCUSSION, LIMITATIONS AND SUGGESTIONS FOR FUTURE DIRECTIONS

The articles we have reviewed are dominated by the focus on higher education. The greater availability of data and the closer relationship researchers have with university institutions (as teaching or research staff) allows more research to be conducted in that setting. On the contrary, Secondary and Primary Education, which are a wider field with a larger population, do not seem to have developed accordingly, as far as the attention they receive from researchers in educational data mining is concerned. Consequently, although there is a large space for research due to older age of students, the variety of subjects, different learning levels of students, etc., the field has not been developed accordingly. Key barriers to further research development are the reduced accessibility of the data, the procedures for approving research by educational authorities, the greater attention paid to students' personal data, and so on.

A wide variety of available algorithms have also been identified that allow for studies to evaluate their findings effectiveness in assessing student performance. Our review has generally shown high levels of accuracy. In only a few cases did we find very low or extremely high levels of accuracy, while no statistically significant differences in accuracy were found between the methods. No improvement was found in using more advanced algorithms either. Decision trees make up the majority of the applied methods with satisfactory levels of accuracy. Naive Bayes, C4.5 and Random Forest algorithms were used more while the KNN algorithm monopolized instance-based methods. K-means was used in only two articles (Clustering). In fewer studies ensemble methods have been applied.

In the literature we studied, student scores were most often used as a predictive variable. Combinations of grades with demographic and academic data were often used. There was no combination that led to greater accuracy, but in each case the average accuracy moved to high levels.

The most common target audience in the articles we studied was the academic community. Most cases were experiments to evaluate the performance of algorithms or suggestions for the use of new algorithms by researchers. Our results showed a satisfactory level of algorithm accuracy. However, the various alternative users of such findings have been almost ignored. In a few cases, a tool was developed or the results of research to improve the education provided by the institutions were applied. The expression of suggestions for the possibility of using the results was more often observed, but without anything specific. So, the goal seems to be biased. However, from a practical point of view, EDM must also provide evidence for the implementation of educational policy or the improvement of the learning process, something that has not been established. In a few cases, the practical application of the research findings has been reported. Such were the use of techniques for early diagnosis of student failure and the improvement of student outcomes through feedback were the only tangible findings.

Overall, this review has identified a rich collection of analysis methods as well as a predominant focus on tertiary education. A limited application of data mining methods has been found to support educational policy-making and institutional decision-making. We believe that the development of research aimed at their own application in the daily teaching process but also in support of decision making at the level of educational policy, should be an alternative. Internal feedback with successful examples of using different algorithms and techniques in the long run, without practical application, can lead to the scientific field withering away. On the contrary, the practical application of the results will expand the scope of research and will be mutually beneficial for the research community and the wider educational community.

This review mainly used descriptive statistics. The diversity of samples reduced the ability of using inductive methods. Extraction of quantitative results from the articles may not be secure as they emanated from different educational environments, different datasets, and different algorithms. We attempted to give the whole picture of a limited range of EDM, which refers to the prediction of academic performance. We used many comprehensive articles after a careful assessment of their content. However, there is plenty of space for scientific research in different areas of EDM, using different types of literature review.

Finally, and most importantly, we believe that broadening of the field beyond higher education, and primarily focusing on the earlier education levels that serve more numerous pupils, in more diverse classes and having a more important impact in their lives and in society as a whole, is the most prominent future direction for this field. It is a direction that can provide new research opportunities, but more importantly a direction that can produce results that affect education and society in a more profound way.

REFERENCES

- Abdulazeez, Y., & Abdulwahab, L. (2019). Application of classification models to predict students' academic performance using classifiers ensemble and synthetic minority over sampling techniques. *Bayero Journal of Pure and Applied Sciences*, 11(2), 142-148. doi: 10.4314/bajopas.v11i2.17
- Al-Noshan, A. A., Al-Hagery, M. A., Al-Hodathi, H. A., & Al-Quraishi, M. S. (2018). Performance evaluation and comparison of classification algorithms for students at Qassim University. *International Journal of Science and Research* 8(11), 1277-1282. doi: 10.21275/ART20202907
- Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, 5 (2019), e01250,. doi: 10.1016/j.heliyon.2019.e01250
- Afeni, B. O., Oloyede, I. A., & Okurinboye, D. (2019). Students' performance prediction using classification algorithms. *Journal of Advances in Mathematics and Computer Science*, 30, 1–9. doi: 10.9734/jamcs/2019/45438
- Agaoglu, M. (2016). Predicting instructor performance using data mining techniques in higher education. *IEEE Access*, 4, 2379–2387. doi: 10.1109/ACCESS.2016.2568756
- Agasisti, T., & Bowers, A. J. (2017). Data analytics and decision making in education: towards the educational data scientist as a key actor in schools and higher education institutions. In Johnes, G., Johnes, J., Agasisti, T., López-Torres, L. (Eds) *Handbook of Contemporary Education Economics*. (pp.184-210). Cheltenham, UK: Edward Elgar Publishing.
- Ahadi, A., Lister, R., Haapala, H., & Vihavainen, A. (2015). Exploring machine learning methods to automatically identify students in need of assistance. In *Proceedings of the 2015 ACM Conference on International Computing Education Research*. (pp. 121 – 130). doi: 10.1145/2787622.2787717
- Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The prediction of students' academic performance using classification data mining techniques. *Applied Mathematical Sciences*, 9(129), 6415–6426. doi: 10.12988/ams.2015.53289
- Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students final gpa using decision trees: A case study. *International Journal of Information and Education Technology*, 6(7), 528–533. doi: 10.7763/ijiet.2016.v6.745

- Alharbi, Z., Cornford, J., Dolder, L., & De La Iglesia, B. (2016). Using data mining techniques to predict students at risk of poor performance. In *Proceedings of 2016 SAI Computing Conference, SAI 2016*, (pp. 523–531). doi: 10.1109/SAI.2016.7556030
- Almasri, A., Celebi, E., & Alkhalwaldeh, R. (2019). MISDataset: Management information systems dataset for predicting undergraduate students' performance. In *Proceedings of the 4th International Conference on Computational Intelligence and Applications, ICCIA 2019*, (pp. 54–57). Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/ICCIA.2019.00017
- Almutairi, S., Shaiba, H., & Bezbradica, M. (2019). Predicting students' academic performance and main behavioral features using data mining techniques. In A. Alfaries, H. Mengash, A. Yasar, & E. Shakshuki (Eds), *Advances in Data Science, Cyber Security and IT Applications*. (pp. 245–259). Cham: Springer International Publishing.
- Al-Saleem, M., Al-Kathiry, N., Al-Osimi, S., & Badr, G. (2015). Mining educational data to predict students' academic performance. In: Perner P. (Eds) *Machine Learning and Data Mining in Pattern Recognition. MLDM 2015. Lecture Notes in Computer Science, vol 9166*. (pp. 403–414). Springer, Cham.
- Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., ... & Olatunji, S. O. (2017). Student performance prediction using Support Vector Machine and K-Nearest Neighbor. *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*. (pp.1-4). IEEE. doi: 10.1109/ccece.2017.7946847
- Altaf, S., Soomro, W., & Rawi, M. I. M. (2019). Student performance prediction using multi-layers artificial neural networks. *ACM Press*. doi: 10.1145/3325917.3325919
- Altujjar, Y., Altamimi, W., Al-Turaiki, I., & Al-Razgan, M (2016). Predicting critical courses affecting students performance: A case study. *Procedia Computer Science*, 82, 65–71. doi: 10.1016/j.procs.2016.04.010
- Aman, F., Rauf, A., Ali, R., Iqbal, F., & Khattak, A. M (2019). A predictive model for predicting students academic performance. In *Proceedings of the 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, (pp. 1-4). doi: 10.1109/IISA.2019.8900760
- Amani, M., & Salama, S. (2019). Discovering performance evaluation features of faculty members using data mining techniques to support decision making. *International Journal of Computer Applications*, 178(49), 25–29. doi: 10.5120/ijca2019919417
- Amaya, Y., Heredia, D & Barrientos, E. (2015). Student dropout predictive model using data mining techniques. *Ieee Latin America Transactions*, 13(9), 3127–3134.
- Amirhajlou, L., Sohrabi, Z., Alebouyeh, M. R., Tavakoli, N., Haghighi, R. Z., Hashemi, A., & Asoodeh, A. (2019). Application of data mining techniques for predicting residents' performance on pre-board examinations: A case study. *Journal of Education and Health Promotion*, 8, (108). Retrieved 11 February 2020 from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6615122/#>
- Amornsinlaphachai, P. (2016). Efficiency of data mining models to predict academic performance and a cooperative learning model. In *Proceedings of the 8th International Conference on Knowledge and Smart Technology, KST 2016*, (pp. 66–71). doi: 10.1109/KST.2016.7440483
- Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015). Preprocessing and analyzing educational data set using X-API for improving student's performance. In *Proceedings of the 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, AEECT 2015*. (pp. 1-5). doi: 10.1109/AEECT.2015.7360581
- Angiani, G., Ferrari, A., Fornacciari, P., Mordonini, M., & Tomaiuolo, M. (2019). Real marks analysis for predicting students' performance. In: Di Mascio T. et al. (Eds) *Methodologies and Intelligent Systems for Technology Enhanced Learning, 8th International Conference. MIS4TEL 2018*, (pp. 37– 44). ChamSpringer International Publishing. doi: 10.1007/978-3-319-98872-6_5

- Anuradha, C., & Velmurugan, T. (2016). Performance evaluation of feature selection algorithms in educational data mining. *International Journal of Data Mining Techniques and Applications*, 5(2), 131–139. doi: 10.20894/IJDMTA.102.005.002.007
- Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students performance using educational data mining. *Computers & Education*, 113, 177–194. doi: 10.1016/j.compedu.2017.05.007
- Athani, S. S., Kodli, S. A., Banavasi, M. N., & Hiremath, P. S. (2017). Student academic performance and social behavior predictor using data mining techniques. In *Proceedings of the International Conference on Computing, Communication and Automation 2017 (ICCCA)*. (pp. 170-174). IEEE. doi: 10.1109/ccaa.2017.8229794
- Baker, R. (2013). Learning, schooling, and data analytics. In: Marilyn M., Redding S., J.Twyman (Eds), *Handbook on innovations in learning*, (pp. 179–190). Center on Innovations in Learning, Temple University, Philadelphia.
- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In: J. A. Larusson, & B. White (Eds), *Learning Analytics: From Research to Practice* (pp. 61–75). New, York, NY: Springer New York. doi: 10.1007/978-1-4614-3305-7_4
- Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM | Journal of Educational Data Mining*, 1(1), 3–17.
- Bergin, S., Mooney, A., Ghent, J., & Quille, K. (2015). Using machine learning techniques to predict introductory programming performance. *International Journal of Computer Science and Software Engineering*, 4(12), 323–328.
- Bhegade, K. B., & Shinde, S. V. (2016). Student performance prediction system with educational data mining. *International Journal of Computer Applications*, 146(5), 32-35. doi: 10.5120/ijca2016910704
- Bowers, A. J. (2017). Quantitative research methods training in education leadership and administration preparation programs as disciplined inquiry for building school improvement capacity. *Journal of Research on Leadership Education*, 12(1), 72–96. doi: /10.7916/D8K93H16
- Brinton, C. G., & Chiang, M. (2015). MOOC performance prediction via clickstream data and social learning networks. In *Proceedings of the IEEE Conference on Computer Communications INFOCOM*, 26, (pp. 2299–2307). doi: 10.1109/INFOCOM.2015.7218617
- Buenaño-Fernández, D., Gil, D., & Luján-Mora, S. (2019). Application of machine learning in predicting performance for computer engineering students: A case study. *Sustainability*, 11, 2833. doi: 10.3390/su11102833
- Buniyamin, N., bin Mat, U., & Arshad, P. M. (2016). Educational data mining for prediction and classification of engineering students achievement. In *Proceedings of the 7th International Conference on Engineering Education, ICEED 2015*, (pp. 49–53). doi: 10.1109/ICEED.2015.7451491
- 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 & Electrical Engineering*, 66, 541–556. doi: 10.1016/J.COMPELECENG.2017.03.005
- Bydžovská, H. (2015). Are collaborative filtering methods suitable for student performance prediction? In *Proceedings of the Portuguese Conference on Artificial Intelligence*. (pp. 425–430). doi: 10.1007/978-3-319-23485-4_42
- Bydžovská, H. (2016). A comparative analysis of techniques for predicting student performance. In *Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016*, 2, (pp. 306–311).
- Canagareddy, D., Subarayadu, K., & Hurbungs, V. (2019). A machine learning model to predict the performance of university students. *Smart and Sustainable Engineering for Next Generation Applications* (pp. 313–322). Springer International Publishing. doi: 10.1007/978-3-030-18240-3_29

- Castro-Wunsch, K., Ahadi, A., & Petersen, A. (2017). Evaluating neural networks as a method for identifying students in need of assistance. In *Proceedings of the Conference on Integrating Technology into Computer Science Education, ITiCSE*, (pp. 111–116). doi: 10.1145/3017680.3017792
- Chango, W., Cerezo, R., & Romero, C. (2019). Predicting academic performance of university students from multi-sources data in blended learning. In *Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems - DATA '19*, (pp. 1–5). New York, New York, USA: ACM Press. doi: 10.1145/3368691.3368694
- Chanlekha, H., & Niramitranon, J. (2018). Student performance prediction model for early-identification of at-risk students in traditional classroom settings. In *Proceedings of the 10th International Conference on Management of Digital EcoSystems - MEDES '18*, (pp. 239–245). doi: 10.1145/3281375.3281403
- Comendador, B. E. V., Rabago, L. W., & Tanguilig, B. T. (2016). An educational model based on Knowledge Discovery in Databases (KDD) to predict learner behavior using classification techniques. In *Proceedings of the IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*. (pp. 1-6). IEEE. doi: 10.1109/icspcc.2016.7753623
- Corrigan, O., Smeaton, A. F., Glynn, M., & Smyth, S. (2015). Using educational analytics to improve test performance. In: Conole G., Kloboučar T., Rensing C., Konert J., Lavoué E. (Eds) *Design for Teaching and Learning in a Networked World. Lecture Notes in Computer Science, vol 9307*. (pp 42–55). Springer, Cham.
- Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students academic failure in introductory programming courses. *Computers in Human Behavior*, 73, 247–256. doi: 10.1016/j.chb.2017.01.047
- Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. In *Proceedings of the 26th International Conference on World Wide Web Companion*. (pp. 415-421). ACM Press. doi: 10.1145/3041021.3054164
- De Sousa, A., Oliveira, C., & Borges, J. (2018). Using academic performance to predict college students dropout: A case study. *Educação e Pesquisa*, 44, 1–17. doi: 10.1590/s1678-4634201844180590
- Del Río, C., & Pineda Insuasti, J. (2016). Predicting academic performance in traditional environments at higher-education institutions using data mining: A review. *Ecós de la Academia*(4). 185-201.
- Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, 94, 335–343. doi: 10.1016/j.jbusres.2018.02.012
- Francis, B. K., & Babu, S. S. (2019). Predicting academic performance of students using a hybrid data mining approach. *Journal of Medical Systems*, 43(6), 162. doi: 10.1007/s10916-019-1295-4
- Gampa, R., & Baynes, A. (2018). Visual analytic workflow to understand students' performance in computer science courses. In *Proceedings of the 2018 IEEE Frontiers in Education Conference (FIE)*, (pp. 1–5). doi: 10.1109/FIE.2018.8658790
- García-Vélez, R. A., López-Nores, M., González-Fernández, G., Robles-Bykbaev, V. E., Wallace, M., Pazos-Arias, J. J., & Gil-Solla, (2019). On Data protection regulations, big data and sledgehammers in higher education. *Applied Sciences*, 9(15), 3084. doi: 10.3390/app9153084
- González-Marcos, A., Olarte-Valentin, R., Ordieres-Meré, J., & Alba-Elias, F. (2019). Predicting students' performance in a virtual experience for project management learning. In *Proceedings of the 11th International Conference on Computer Supported Education*, (pp. 665–673). doi: 10.5220/0007843506650673
- Govindasamy, K., & Velmurugan, T. (2017). A study on classification and clustering data mining algorithms based on students academic performance prediction. *International Journal Of Control Theory And Applications*, 10(23), 147–160.

- Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015). Predicting students performance in educational data mining. In *2015 International Symposium on Educational Technology (ISET)*, (pp. 125-128). IEEE. doi: 10.1109/ISET.2015.33
- Hamsa, H., Indiradevi, S., & Kizhakkethottam, J. (2016). Student academic performance prediction model using decision tree and fuzzy genetic algorithm. *Procedia Technology*, 25, 326-332. doi: 10.1016/j.protcy.2016.08.114
- Hasheminejad, S. M., & Sarvmili, M. (2018). S3PSO: Students' performance prediction based on particle swarm optimization. *Journal of AI and Data Mining* 7(1), 77-96. doi: 10.22044/jadm.2018.5506.1662
- Hassan, H., Anuar, S., & Ahmad, N. B. (2019). Students' performance prediction model using meta-classifier approach. In *Proceedings of the International Conference on Engineering Applications of Neural Networks*. (pp. 221-231). Springer, Cham. doi: 10.1007/978-3-030-20257-6_19
- He, J., Bailey, J., Rubinstein, B. I., & Zhang, R. (2015, February). Identifying at-risk students in massive open online courses. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*. Retrieved 5 February 2020 from: <https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9696/9460>
- Huang, A. Y., Lu, O. H., Huang, J. C., Yin, C. J., & Yang, S. J. (2019). Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. *Interactive Learning Environments*, 28(2), 206-230. doi: 10.1080/10494820.2019.1636086
- Hussain, M., Hussain, S., Zhang, W., Zhu, W., Theodorou, P., & Abidi, S. M. R.. (2018). *Mining moodle data to detect the inactive and low-performance students during the moodle course*. Presented at the 2nd International Conference on Big Data Research - ICBDR 2018 (pp. 133-140). ACM Press. doi: 10.1145/3291801.3291828
- Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, 9(2), 447-459. doi: 10.11591/ijeecs.v9.i2.pp447-459
- Hussain, S., Muhsion, Z. F., Salal, Y. K., Theodoru, P., KurtoÄyllu, F., & Hazarika, G. C. (2019). Prediction model on student performance based on internal assessment using deep learning. *International Journal of Emerging Technologies in Learning (IJET)*, 14(8), 4-22. doi: 10.3991/ijet.v14i08.10001
- Iam-On, N., & Boongoen, T. (2015). Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics*, 8, 497-510. doi: 10.1007/s13042-015-0341-x
- Inyang, U. G., Umoh, U. A., Nnaemeka, I. C., & Robinson, S. A. (2019). Unsupervised characterization and visualization of students' academic performance features. *Computer and Information Science*, 12(2), 103-116. doi: 10.5539/cis.v12n2p103
- Hridoy, S. A. A., Ekram, M. T., Islam, M. S., Ahmed, F., & Rahman, R. M. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics*, 2(1), 1-25. doi: 10.1186/s40165-014-0010-2
- Kamal, P., & Ahuja, S. (2019). An ensemble-based model for prediction of academic performance of students in undergrad professional course. *Journal of Engineering, Design and Technology*, 17(4), 769-781. doi: 10.1108/JEDT-11-2018-0204
- Kasthuriarachchi, K. T., & Liyanage, S. R. (2019). Predicting students' academic performance using utility based educational data mining. In: Hung J., Yen N., Hui L. (Eds) *Frontier Computing. FC 2018. Lecture Notes in Electrical Engineering*, vol 542. (pp. 29-39) Springer, Singapore. doi: 10.1007/978-981-13-3648-5_4
- Kaur, G., & Singh, W. (2016). Prediction of student performance using weka tool. *International Journal of Engineering Sciences*, 17, 8-16. Retrieved 10 February 2020 from: <http://ijoes.vidyapublications.com/paper/Vol17/02-Vol17.pdf>

- Ketui, N., Wisomka, W., & Homjun, K. (2019). *Using classification data mining techniques for students performance prediction*. Presented at the Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON) 2019. IEEE. (pp. 359-363). doi: 10.1109/ecti-ncon.2019.8692227
- Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2010). Systematic literature reviews in software engineering—a tertiary study. *Information and software technology*, 52(6), 792–805. doi: 10.1016/j.infsof.2010.03.006
- Kiu, C. (2018). *Data mining analysis on student's academic performance through exploration of student's background and social activities*. Presented at the 4th International Conference on Advances in Computing, Communication and Automation, ICACCA 2018, (pp. 1–5). doi: 10.1109/ICACCAF.2018.8776809
- Klusener, M., & Fortenbacher, A. (2015). *Predicting students' success based on forum activities in MOOCs*. Presented at the IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2015, 2, (pp. 925–928). doi: 10.1109/IDAACS.2015.7341439
- Knowles, J. (2015). Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining*, 7(3), 18–67.
- Koprinska, I., Stretton, J., & Yacef, K. (2015, June). Predicting student performance from multiple data sources. In *International Conference on Artificial Intelligence in Education* (pp. 678-681). Springer, Cham.
- Koprinska, I., Stretton, J., & Yacef, K. (2015). *Students at risk: Detection and remediation*. Presented at the 8th International Conference on Educational Data Mining. (pp. 512 – 515).
- Kostopoulos, G., Kotsiantis, S., & Pintelas, P. (2015). *Predicting student performance in distance higher education using semi-supervised techniques*. In: Bellatreche L., Manolopoulos Y. (Eds) *Model and Data Engineering. Lecture Notes in Computer Science*, vol 9344. (pp. 259–270). Springer, Cham doi: 10.1007/978-3-319-23781-7_21
- Liang, J., Li, C., & Zheng, L. (2016). *Machine learning application in MOOCs: Dropout prediction*. Presented at the 11th International Conference on Computer Science & Education (ICCSE) 2016, IEEE. (pp. 52–57). doi: 10.1109/ICCSE.2016.7581554
- Livieris, I., Mikropoulos, T., & Pintelas, P. (2016). A decision support system for predicting students' performance. *Themes in Science and Technology Education*, 9(1), 43–57.
- Livieris, I. E., Drakopoulou, K., Mikropoulos, T. A., Tampakas, V., & Pintelas, P. (2018). An ensemble-based semi-supervised approach for predicting students' performance. In: Mikropoulos T. (Eds) *Research on e-Learning and ICT in Education*. (pp. 25-42). Springer Cham. doi: 10.1007/978-3-319-95059-4
- Livieris, I. E., Drakopoulou, K., Tampakas, V. T., Mikropoulos, T. A., & Pintelas, P (2019). Predicting secondary school students' performance utilizing a semi-supervised learning approach. *Journal of Educational Computing Research*, 57(2), 448–470. doi: 10.1177/0735633117752614
- Livieris, I. E., Kotsilieris, T., Tampakas, V., & Pintelas, P. (2019). Improving the evaluation process of students' performance utilizing a decision support software. *Neural Computing and Applications*, 31, 1683–1694. doi: 10.1007/s00521-018-3756-y
- Livieris, I. E., Tampakas, V., Kiriakidou, N., Mikropoulos, T., & Pintelas, P. (2019). Forecasting students' performance using an ensemble ssl algorithm. *Communications in Computer and Information Science* (pp. 566–581). Springer. doi: 10.1007/978-3-030-20954-4_43
- Guarín, C. E. L., Guzmán, E. L., & González, F. A.. (2015). A model to predict low academic performance at a specific enrollment using data mining. *Revista Iberoamericana de Tecnologías del Aprendizaje*, 10(3), 119–125. doi: 10.1109/RITA.2015.2452632
- Le Mai, T., Do, P. T., Chung, M. T., & Thoai, N. (2019). An apache spark-based platform for predicting the performance of undergraduate students. In *2019 IEEE 21st International Conference on High*

Performance Computing and Communications; IEEE (pp. 191-199).. doi: 10.1109/hpcc/smartycity/dss.2019.00041

- Lu, O. H., Huang, A. Y., Huang, J. C., Lin, A. J., Ogata, H., & Yang, S. J. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. *Educational Technology and Society*, 21(2), 220–232.
- Luo, J., Sorour, S. E., Goda, K., & Mine, T. (2015). Predicting student grade based on free-style comments using word2vec and ann by considering prediction results obtained in consecutive lessons. *International Educational Data Mining Society*. (pp. 396–399).
- Mahboob, T., Irfan, S., & Karamat, A.. (2017). *A machine learning approach for student assessment in E-learning using Quinlan's C4.5, Naive Bayes and Random Forest algorithms*. Presented at the 19th International Multi-Topic Conference, INMIC 2016. (pp. 1 -8). doi: 10.1109/INMIC.2016.7840094
- Martins, M. P., Migueis, V. L., & Fonseca, D. S. B. (2018, June). Educational data mining: A literature review. In *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE. (pp. 1-6). doi: 10.23919/CISTI.2018.8399281
- Meedech, P., Iam-On, N., & Boongoen, T (2016). Prediction of student dropout using personal profile and data mining approach. New York. In: Lavanganananda K., Phon-Amnuaisuk S., Engchuan W., Chan J. (Eds) *Intelligent and Evolutionary Systems. Proceedings in Adaptation, Learning and Optimization*, vol 5. Springer, Cham (pp. 143-155). doi: 10.1007/978-3-319-27000-5_12
- Mehboob, B., Liaqat, R., Abbas, N. (2017). Student performance prediction and risk analysis by using data mining approach. *Journal of Intelligent Computing*, 14(7), 69–76.
- Merchan Rubiano, S. M., & Duarte Garcia, J. A. (2016). Analysis of data mining techniques for constructing a predictive model for academic performance. *IEEE Latin America Transactions*, 14 (6), 2783–2788. doi: 10.1109/TLA.2016.7555255
- Mhetre, V., & Nagar, M. (2017). *Classification based data mining algorithms to predict slow, average and fast learners in educational system using WEKA*. Presented at the International Conference on Computing Methodologies and Communication (ICCMC). IEEE. (pp. 475-479). doi: 10.1109/iccmc.2017.8282735
- Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, 115, 36–51. doi: 10.1016/j.dss.2018.09.001
- Moscoso-Zea, O., Saa, P., & Luján-Mora, S. (2019). Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining. *Australasian Journal of Engineering Education*, 24(1), 4–13. doi: 10.1080/22054952.2019.1601063
- Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International Journal of Modern Education and Computer Science*, 8(11), (pp. 36–42). doi: 10.5815/ijmeecs.2016.11.05
- Mundada, O. (2016). Mining educational data from students' management system. *International Journal of Advanced Research in Computer Science*, 7(3), 244–249.
- Namomsa, G., & Sharma, D. P. (2018). Comparative analytics and predictive modeling of student performance through data mining techniques. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3395126
- Naser, S. S., Zaqout, I. S., Abu Ghosh, M., Atallah, R. R., & Alajrami, E. (2015). Predicting student performance using artificial neural network: in the faculty of engineering and information technology. *International Journal of Hybrid Information Technology*, 8(2), 221–228. doi: 10.14257/ijhit.2015.8.2.20
- Navamani, J. M., & Kannammal, A. (2015). Predicting performance of schools by applying data mining techniques on public examination results. *Research Journal of Applied Sciences, Engineering and Technology*, 9(4), 262–271. doi: 10.19026/rjaset.9.1403

- Nemomsa, G., & Mulugeta, A. (2018). Comparative analytics and predictive modeling of student performance through data mining techniques. SSRN. doi: 10.2139/ssrn.3382511
- Nuñez, H., Maldonado, G., & Astudillo, C. A. (2018). *Semi-supervised regression based on tree SOMs for predicting students performance*. Presented at the 9th International Conference on Pattern Recognition Systems (ICPRS 2018). Institution of Engineering and Technology. (pp. 12-19). doi: 10.1049/cp.2018.1288
- Nust, C. E. (2017). Student performance prediction and risk analysis by using data mining approach. *Journal of Intelligent Computing Volume*, 8(2), 49-57.
- Okubo, F., Yamashita, T., Shimada, A., & Ogata, H. (2017). *A neural network approach for students' performance prediction*. Presented at the the Seventh International Learning Analytics & Knowledge Conference on - LAK '17, (pp. 598-599). doi: 10.1145/3027385.3029479
- Pai, M., McCulloch, M., Gorman, J. D., Pai, N., Enanoria, W., Kennedy, G., ... & Colford, J. J. (2004). Systematic reviews and meta-analyses: an illustrated, step-by-step guide. *The National medical journal of India*, 17 (2), 86-95.
- Pandey, M., & Taruna, S. (2016). Towards the integration of multiple classifier pertaining to the Student's performance prediction. *Perspectives in Science*, 8, 364-366. doi: 10.1016/j.pisc.2016.04.076
- Papamitsiou, Z., & Economides, A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, 17(4), 49-64.
- Parmar, K., Vaghela, D., & Sharma, P. (2015). *Performance prediction of students using distributed data mining*. Presented at the: International Conference on Innovations in Information, Embedded and Communication Systems 2015 (ICIIECS), (pp. 1-5). doi: 10.1109/ICIIECS.2015.7192860
- Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 4(1), 1432-1462. doi: 10.1016/j.eswa.2013.08.042
- Polyzou, A., & Karypis, G. (2018). Feature extraction for classifying students based on their academic performance. *IEEE Transactions on Learning Technologies*, 12(2), 237-248. doi: 10.1109/TLT.2019.2913358
- Pradeep, A., Das, S., & Kizhekkethottam, J. J. (2015). *Students dropout factor prediction using EDM techniques*. Presented at the IEEE International Conference on Soft-Computing and Network Security, ICSNS 2015. (pp. 1-7). doi: 10.1109/ICSNS.2015.7292372
- Pristyanto, Y., Pratama, I., & Nugraha, A. F. (2018). *Data level approach for imbalanced class handling on educational data mining multiclass classification*. Presented at the International Conference on Information and Communications Technology, ICOIACT 2018, (pp. 310-314). doi: 10.1109/ICOIACT.2018.8350792
- Rivas, A., Fraile, J. M., Chamoso, P., González-Briones, A., Rodríguez, S., & Corchado, J. M. (2019). Students performance analysis based on machine learning techniques. In: Uden L., Liberona D., Sanchez G., Rodríguez-González S. (Eds) *Learning Technology for Education Challenges. LTEC 2019. Communications in Computer and Information Science*. (pp. 428-438). Springer International Publishing. doi: 10.1007/978-3-030-20798-4_37
- Rodrigues, R. L., Ramos, J. L. C., Silva, J. C. S., Dourado, R. A., & Gomes, A. S. (2019). Forecasting students' performance through self-regulated learning behavioral analysis. *International Journal of Distance Education Technologies*, 17, 52-74. doi: 10.4018/ijdet.2019070104.
- Rodriguez, L. G., Rodriguez, C. A., Niguidula, J. D., & Calibo, D. I. (2019). Analysis of students performance: Input to program enhancement of students in computing. *International Journal of Information and Education Technology*, 9(6), 429-432. doi: 10.18178/ijiet.2019.9.6.1240
- Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33 (1), 135-146. doi: 10.1016/j.eswa.2006.04.005

- Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40 (6), 601–618. doi: 10.1109/TSMCC.2010.2053532
- Ruby, J., & David, K. (2015). Analysis of influencing factors in predicting students performance using MLP - A comparative study. *International Journal of Innovative Research in Computer and Communication Engineering (An ISO Certified Organization)*, 3(2), 1085–1092. doi: 10.15680/ijircce.2015.0302070
- S. Abu-Oda, G., & M. El-Halees, A. (2015). Data mining in higher education : University Student dropout case study. *International Journal of Data Mining & Knowledge Management Process*, 5(1), 15–27. doi: 10.5121/ijdkp.2015.5102
- Saa, A. A., Al-Emran, M., & Shaalan, K. (2019). *Mining student information system records to predict students' academic performance*. Presented at the International conference on advanced machine learning technologies and applications. (pp. 229–239). Springer Cham. doi: 10.1007/978-3-030-14118-9_23
- Kalyani, B. S., Harisha, D., RamyaKrishna, V., & Manne, S.. (2019). Evaluation of students performance using neural networks. In: Pandian, A. P., Ntalianis, K., & Palanisamy, R. (Eds). *Intelligent Computing, Information and Control Systems. Advances in Intelligent Systems and Computing*. (pp. 499 - 505). Springer. doi: 10.1007/978-3-030-30465-2_55
- Santoso, L. W., & Yulia. (2019). The analysis of student performance using data mining. In: Bhatia S., Tiwari S., Mishra K., Trivedi M. (Eds) *Advances in Computer Communication and Computational Sciences* (pp. 559-573). Springer. doi: 10.1007/978-981-13-6861-5_48
- Sara, N. B., Halland, R., Igel, C., & Alstrup, S. (2015). *High-school dropout prediction using machine learning High-School Dropout Prediction Using Machine Learning : A Danish Large-scale Study*. Presented at the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, (pp. 319–324). Bruges.
- Saravanan, M., & Jyothi, V. L. (2019). A novel approach for performance and estimation of students by using hybrid model. *IIOAB Journal*, 10(2), 9–12.
- Satyanarayana, A., & Nuckowski, M. (2016). *Data mining using ensemble classifiers for improved prediction of student academic performance data mining using ensemble classifiers for improved prediction of student academic performance*. City University of New York (CUNY), New York. Retrieved 10 February 2020 from: https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1078&context=ny_pubs
- Shrivasa, A. K., & Tiwari, P. (2017). Comparative analysis of models for student performance with data mining tools. *International Journal of Computer Trends and Technology*, 46(1), 42–46. doi: 10.14445/22312803/ijctt-v46p109
- Singh, J., & Kaur, E. H. (2018). Implementation of classification based algorithms to prediction of students performance using educational system. *International Journal of Engineering Science and Computing*, 8, 19233–19238.
- Sivakumar, S., & Selvaraj, R. (2018). Predictive modeling of students performance through the enhanced decision tree. *Advances in Electronics, Communication and Computing*, 21-36. doi: 10.1007/978-981-10-4765-7_3
- Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology*,9(4), 1-5. doi: 10.17485/ijst/2016/v9i4/87032
- Son, L. H., & Fujita, H. (2019). Neural-fuzzy with representative sets for prediction of student performance. *Applied Intelligence*, 49(1), 172–187. doi: 10.1007/s10489-018-1262-7
- Sorour, S. E., Goda, K., & Mine, T. (2015). *Estimation of Student performance by considering consecutive lessons*. Presented at the 4th International Congress on Advanced Applied Informatics, Okayama, 2015, (pp. 121-126), doi: 10.1109/IIAI-AAI.2015.170

- Strecht, P., Cruz, L., Soares, C., & Mendes-Moreira, J. (2015). A comparative study of classification and regression algorithms for modelling students' academic performance. In *Proceedings of the 8th International Conference on Educational Data Mining*, 392–395. Retrieved 9 February 2020 from: <http://www.educationaldatamining.org/EDM2015/proceedings/short392-395.pdf>.
- Sukhija, K., Jindal, M., & Aggarwal, N. (2015). *The recent state of educational data mining: A survey and future visions*. Presented at the 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE), (pp. 354–359). doi: 10.1109/MITE.2015.7375344
- Sumitha, R., Vinothkumar, E. S., & Scholar, P. G. (2016). Prediction of students outcome using data mining techniques. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, 2(6), 132–139.
- Tang, J. K., Xie, H., & Wong, T. L. (2015). A big data framework for early identification of dropout students In: J. Lam, K. K. Ng, S. K. Cheung, T. L. Wong, K. C. Li, & F. L. Wang (Eds) , *Communications in Computer and Information Science*, (pp. 127–132). Springer Berlin Heidelberg. doi: 10.1007/978-3-662-48978-9_12
- Tasnim, N., Paul, M. K., & Sattar, A. S. (2019). *Performance analysis of different decision tree based methods for identifying drop out students*. Presented at the 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), (1–6). IEEE. doi: 10.1109/ICASERT.2019.8934518
- Tegegne, A. K., & Alemu, T. A. (2018). Educational data mining for students' academic performance analysis in selected Ethiopian universities. *Information Impact: Journal of Information and Knowledge Management*, 9(2), 1-15. doi: 10.4314/ijikm.v9i2.1
- Thakar, P., Anil Mehta, A. (2015). Performance analysis and prediction in educational data mining: A research travelogue. *International Journal of Computer Applications*, 110(15). 60-68.
- Timaran-Pereira, R., Hidalgo, A., Caicedo, J., & Benavides, J. (2018). Discovering factors associated with academic performance of high school students. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*. (pp. 188-197). Association for the Advancement of Computing in Education (AACE).
- Ulloa-Cazarez, R. L., López-Martín, C., Abran, A., & Yáñez-Márquez, C. (2018). Prediction of online students performance by means of genetic programming. *Applied Artificial Intelligence*, 32(9), 858–881. doi: 10.1080/08839514.2018.1508839
- Widyaningsih, Y., Fitriani, N., & Sarwinda, D. (2019). *A semi-supervised learning approach for predicting student ' s performance : First-year*. Presented at the 12th International Conference on Information & Communication Technology and System (ICTS), (pp. 291–295)
- Yehuala, M. A. (2015). Application of data mining techniques for student success and failure prediction the case of Debre Markos University. *International Journal of Scientific & Technology Research*,4 (4), 91–94.
- Zhang, X., Sun, G., Pan, Y., Sun, H., He, Y., & Tan, J. (2018). Students performance modeling based on behavior pattern. *Journal of Ambient Intelligence and Humanized Computing*, 9(5), 1659–1670. doi: 10.1007/s12652-018-0864-6
- Zhou, Q., Zheng, Y., & Mou,. (2015). Predicting students' performance of an offline course from their online behaviors. In *2015 Fifth International Conference on Digital Information and Communication Technology and its Applications (DICTAP)*, (pp. 70–73). doi: 10.1109/DICTAP.2015.7113173