JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 18, No. 2 (2022), pp. 58-69

Educational data mining, student academic performance prediction, prediction methods, algorithms and tools: an overview of reviews

Chaka Chaka[a,1]

[a]*University of South Africa, Dept. of English Studies – Pretoria (South Africa)*

(submitted: 1/11/2021; accepted: 30/8/2022; published: 31/8/2022)

Abstract

This overview study set out to compare and synthesise the findings of review studies conducted on predicting student academic performance (SAP) in higher education using educational data mining (EDM) methods, EDM algorithms, and EDM tools from 2013 to September 2021. It conducted multiple searches for suitable and relevant peer-reviewed articles on two online search engines, on nine online databases, and on two online academic social networks. It, then, selected 33 eligible articles from 2,500 articles. Some of the findings of this overview study are worth mentioning. First, only 3 studies explicitly stated their precise sample sizes, and only 5 studies explicitly mentioned their subject areas with maths and science, and computer science and engineering as the four most mentioned subject areas. Second, 20 review studies had purposes related to either EDM techniques, EDM methods, EDM models, or EDM algorithms employed to predict SAP and student success in the higher education sector. Third, there are six commonly used typologies of input variables reported by 33 review studies, of which student demographics was the most commonly utilised variable for predicting SAP. Fourth and last, seven common EDM algorithms employed for predicting SAP were identified, of which Decision Tree emerged both as the most used algorithm and as the algorithm with the highest prediction accuracy rate for predicting SAP.

KEYWORDS: Overview, Student Academic Performance, Educational Data Mining, Methods, Algorithms.

DOI
https://doi.org/10.20368/1971-8829/1135578

CITE AS
Chaka, C. (2022). Educational data mining, student academic performance prediction, prediction methods, algorithms and tools: an overview of reviews. *Journal od e-Learning and Knowledge Society*, *18*(2), 58-69.
https://doi.org/10.20368/1971-8829/1135578

1. Introduction

The last few years have witnessed an exponential increase in review studies exploring educational data mining (EDM) methods, algorithms, and tools for predicting student academic performance (SAP) (Khasanah, 2018; Saa et al., 2019; Shahiri et al., 2015). This is the case for diverse disciplinary fields, even though fields such as computer science and engineering seem to have conducted more such studies than others (Ashenafi, 2017). Most EDM review studies on predicting SAP have been conducted as either reviews (Ameen et al., 2019; Cui et al., 2019; Del Río & Insuasti, 2016; Durga & Thangakumar, 2019; Moreno-Marcos et al., 2019; Muttathil & Rahman, 2016; Shahiri et al., 2015); literature reviews (Alyahyan & Düştegör, 2020; Manjarres et al., 2018; Saqr, 2018); systematic literature reviews (Alban & Mauricio, 2019; Liz-Domínguez et al., 2019; Namoun & Alshanqiti, 2021; Papamitsiou & Economides, 2014); systematic reviews (Agrusti et al., 2019; Alamri & Alharbi, 2021; Aydogdu, 2020; López-Zambrano et al., 2021; Zulkifli et al., 2019); review syntheses (Aldowah et al., 2019); or surveys (Alturki et al., 2020; Ganesh & Christy, 2015; Jindal & Borah, 2013). While these review study types are not exhaustive, they represent a broad spectrum of the types of review studies that the current paper was able to locate.

2. Contextualising issues

This paper uses an overview of reviews in the same sense as a review of reviews. In an overview of reviews (hereafter an overview or an overview study), review studies or aspects featuring in review studies become key units or foci of analysis as opposed to aspects of

[1] corresponding author - email: chakachaka8@gmail.com

primary studies (Polanin et al., 2016). There are different terms used to refer to a review of reviews. These include review of reviews, second-order review, umbrella review, tertiary review, meta-meta-analysis, synthesis of meta-analysis, synthesis of systematic reviews, summary of systematic reviews, or systematic review of systematic reviews (Grant & Booth, 2009; Moonsamy et al., 2021; Pieper et al., 2012). These terms constitute typologies of overviews. These typologies reflect the roles played by the respective overviews and the purposes these overviews are meant to serve.

Benefits of utilising overviews are:

- retrieving, identifying, assessing and integrating findings from several review studies leveraging previous research syntheses;
- aggregating the evidence provided by multiple reviews or contrasting multiple treatments on the same topic; and
- identifying a gap in existing reviews (Pieper et al., 2012; Polanin et al., 2016).

3. Literature review related to predicting student academic performance using EDM techniques

Student academic performance (SAP) is a crucial construct employed to determine student academic success at different educational levels (Khanna et al., 2016; Papadogiannis et al., 2020; Shahiri et al., 2015). Even though it has multiple definitions, at a basic level, SAP is the performance that students display in their academic tasks (e.g., assignments, tests and examinations). It is often reflected in students' past cumulative grade point average (CGPA)/grade point average (GPA) in a previous semester and in students' expected GPA in the existing semester. If the term *performance* is disaggregated from the phrase *student academic performance*, it embodies achievement in relation to assignments and courses, continuous progress in programmes, and a successful completion of programmes (Hellas et al., 2018; Khasanah, 2018). Moreover, it entails persistence, retention, progression, wastage, and success or progress (Hamoud et al., 2017). In this sense, SAP should be seen in the same way as student academic achievement (Alyahyan et al., 2020). However, SAP is a complex construct, and in this regard, there are multiple factors that impact on and affect it. These include the historical academic performance and the socio-economic background of students.

In this context, some of the factors (also known as attributes) employed to predict SAP are: academic factors (historical and current); student demographics; socio-economics factors; psychological factors; student e-learning activities; student environments; and extra-curricular activities (Kumar & Salal, 2019). The superordinate factors listed in the preceding set are often utilised to predict SAP by most scholars (Alturki et al., 2020; Khasanah, 2018). These superordinate factors are further categorised into specific subordinate factors with the former serving as input variables or performance features, and with the latter serving as output variables or performance metrics [18]. Nonetheless, at times there are overlaps between the superordinate and subordinate factors as certain scholars tend to conflate them (Alyahyan & Düştegör, 2020; Ashenafi, 2017; Hellas et al., 2018; Kumar & Salal, 2019).

Moreover, certain methods (or tasks) such as association rule mining, clustering, classification and regression are used for building models for predicting SAP. Such methods are at times referred to as techniques (Aldowah et al., 2019; Hellas et al., 2018), while Saa et al. (2019) call them EDM approaches. In this way, classification tends to be the predominantly used method. Furthermore, there are algorithms that are employed to predict SAP. Among them are Artificial Neural Network (ANN), Bayesian Network (BN), Decision Tree (DT), K-Nearest Neighbours (K-NN), K-Means; Naïve Bayesian classifiers, Neural Network (NN), and Support Vector Machine (SVM) (Alamri & Alharbi, 2021; Ashenafi, 2017; López-Zambrano et al., 2021; Namoun & Alshanqiti, 2021).

In certain instances, these algorithms are referred to as EDM techniques (Ashenafi, 2017), or as tasks or as methods (Alturki et al., 2020). The choice of prediction algorithms is determined by SAP outcomes to be predicted. For instance, classification algorithms such as DT, NN and NB classifiers are commonly used for predicting a binary outcome like pass/fail at a certain degree of probability (Alamri & Alharbi, 2021; Ashenafi, 2017; Shahiri et al., 2015). By contrast, SVM and linear regression are often employed for predicting numerical scores (Ashenafi, 2017). Furthermore, some of the tools belonging to software programmes such as WEKA, RapidMiner, MATLAB, KNIME, Apache Mahout, Rattle GUI are used for predicting SAP. Of these, WEKA appears to be the frequently used tool (Alyahyan & Düştegör, 2020; Alturki et al., 2020; Khasanah, 2018; Kumar & Salal, 2019).

4. Purpose of the study

The purpose of this paper was to compare and synthesise findings of review studies conducted on predicting SAP in higher education through utilising EDM methods, algorithms, and tools from 2013 to September 2021. The major focus was on review studies related to the higher education sector. The following served as research questions (RQs) for this study:

- RQ1: What are the primary purposes of the review studies investigated in this overview?
- RQ2: What common input (predictor) and common output (target) variables do these review studies employ to predict SAP?

- RQ3: What common educational data mining (EDM) techniques (or methods) and algorithms do they employ in predicting SAP?
- RQ4: What algorithms are reported to have the highest prediction accuracy for SAP?
- RQ5: What common EDM tools do these studies employ in predicting SAP?
- RQ6: What are the key results of these review studies?

5. Literature search strategy

The search strategy for relevant review studies was conducted online from March 2020 to September 2021, and started by locating search engines, databases, and academic social networking sites. Subsequently, two online search engines (Google and Bing), nine online databases (Google Scholar, Microsoft Academic, Semantic Scholar, IEEE Xplore, ERIC, ScienceDirect, Emerald; JSTOR, SpringerLink), and two online academic social networks (ResearchGate and Academia.edu), were identified (Figure 1). Search strings were arranged into super- and sub-strings in keeping with the major area of focus of the overview: predicting SAP through using EDM methods, algorithms, and tools. These search strings consisted of the following keywords: predicting student academic performance; educational data mining techniques; educational data mining algorithms; and educational data mining software tools. To ensure that a wide range of review studies on the major focus area of this overview was covered in all the search combinations, two commonly used Boolean operators, *AND* and *OR*, together with parentheses and double quotation marks (where necessary), were employed in the search strategy. Examples of these search combinations were as follows:

- predicting student academic performance AND educational data mining techniques AND educational data mining algorithms AND educational data mining software tools
- predicting student academic performance OR educational data mining techniques OR educational data mining algorithms OR educational data mining software tools.

In certain instances, the word, *techniques*, was replaced with *methods* and *tasks*. The afore-said keyword combinations, together with their relevant iterations, were queried in the three search engines, in the nine online databases, and in the two online academic social networking sites mentioned earlier. Moreover, dependency and snowball search strategies were employed based on the bibliographies of the journal articles obtained from the three sets of online search platforms.

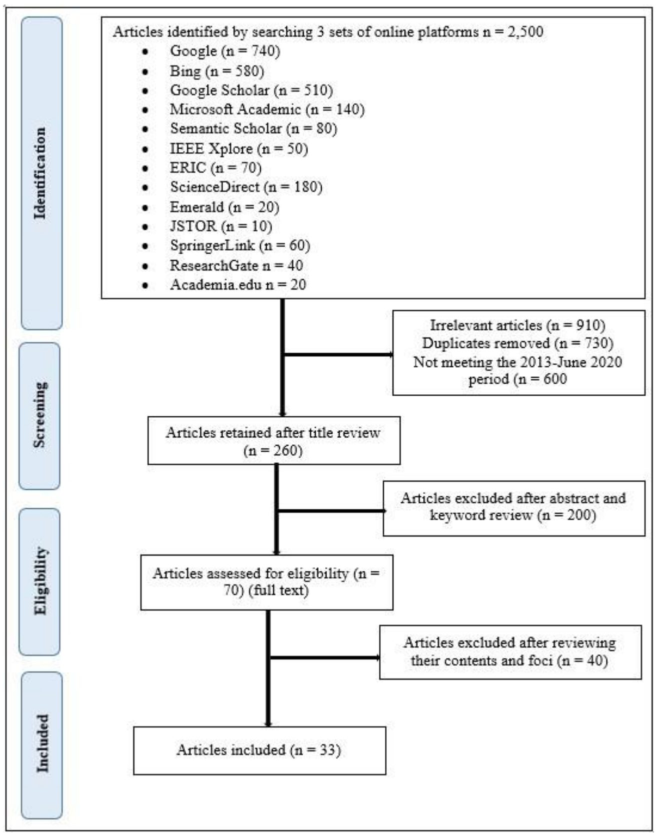


Figure 1 - The PRISMA flow chart and the online search platforms.

5.1 Eligibility criteria and selection of studies

The criteria for including and excluding review studies are as listed below. They were formulated to respond to the major focus area of the current overview.

- review studies focusing on predicting SAP using EDM methods (techniques or tasks), algorithms and tools;
- focus on higher education;
- review studies published between 2013 and September 2021;
- review studies published in peer-reviewed journals and by (internationally) recognised conference organisations;
- mention of a specific years/duration covered (e.g., 2010 to 2015); and
- review studies published in English.

Review studies were identified and selected by following a four-phase selection process informed by the PRISMA approach as illustrated in Figure 1. One of the key aspects of this approach is to ascertain that there is clarity and transparency in the search and selection processes ((Moher et al., 2009). The first phase involved screening articles, which were obtained from the three sets of online search platforms by querying a combination of search strings mentioned earlier. This phase yielded 2,500 articles. The second phase entailed screening these articles by reviewing their titles. This resulted in 260 articles being retained. Thereafter, the third phase was conducted during which 200 irrelevant and duplicate articles were eliminated by reviewing their abstracts and keywords. In the fourth phase, 30 irrelevant articles were identified and excluded after review their contents and foci, resulting in 33 full-text articles judged as relevant being retained from 40 articles. These 33 articles served as the major source of data sets for the current overview.

5.2 Data extraction, coding and inter-rater reliability

Data sets, based on the purpose and on the major focus area of the overview, were extracted from 33 full-text articles mentioned above. A coding scheme consisting of categories based on 14 specific features of the major focus area (Appendix A) was developed. Examples of these categories are: total sample size; purpose of review; input variables; output variables; and EDM techniques. Raters used this coding scheme to extract data from the 33 articles, code them, and match them to each of these categories. To ensure data extraction and data coding consistency, three raters extracted and coded data. The coding protocol used was based on Miles and Huberman's (1994) inter-rater reliability (IRR), which employs the following formula:

$$reliability = \frac{number\ of\ agreements}{number\ of\ agreements + disagreements}$$

In keeping with this formula, the three raters had a mean IRR of 77% agreement for all the data they had coded for the 14 categories. An IRR of 77% agreement is deemed to be sufficiently reliable (Miles & Huberman, 1994).

5.3 Data analysis

Two related and complimentary techniques were used to analyse data sets: content analysis and thematic analysis. The choice of these two analytic approaches was informed by the types of data sets extracted from the 33 articles. Content analysis lent itself well to quantitatively representing categories and themes extracted from the data, while thematic analysis was employed to qualitatively present these categories and themes (Vaismoradi & Snelgrove, 2019).

6. Findings

he findings presented in this section of the overview are grounded on the data extracted from the 33 full-text articles and are informed by the manner in which the extracted data were codified as highlighted in the relevant section above. Additionally, the findings respond to the six research questions stated earlier.

6.1 A Panoramic view of the thirty-three review studies

Of the 33 review studies investigated, 9 were systematic reviews; 8 were reviews; 5 were systematic reviews; 4 were surveys; 3 were literature reviews; and the last 4 were a meta-analysis, a critical review, a comparative analysis, and a review and synthesis, apiece (Figure 2). In all, there were nine different types of reviews, with systematic reviews as a typology constituting the most of these review studies.

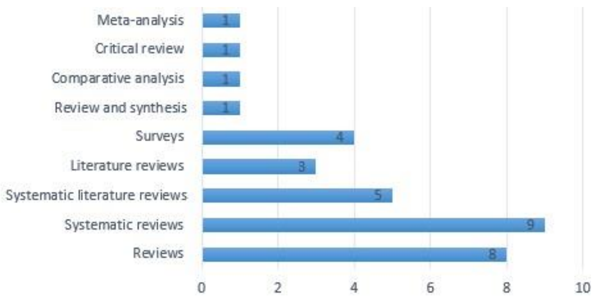


Figure 2 - Types of review studies reviewed.

Additionally, these 33 review studies had their authors from diverse albeit, in some cases, the same countries of origin. For instance, on the one hand, as depicted in Figure 3, 6 reviews were written by authors based in India, while 4 studies and 3 studies were written by authors from Saudi Arabia and Spain, respectively. On the other hand, 2 reviews had authors from Greece, Malaysia, Italy, and Germany, each. The remaining 12 reviews were written by authors from either single, dual, or multiple countries.

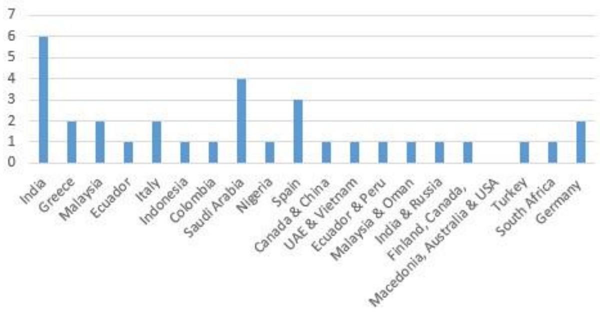
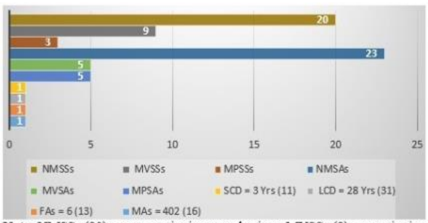


Figure 3 - Authors' and review studies' countries of origin.

Even though the 33 review studies were published between 2013 and September 2021, the aggregate time duration covered by these studies spanned 29 years (1992-2021) (Table 1). The study with the longest duration (longest time span) is review study 31, which covered a 28-year duration (1992-November 2020) (see Figure 4 and Appendix A). This study contrasts with review study 11, whose duration is 3 years (2007-2010). The study that had the most articles is review study 16, which reviewed 402 articles. Its converse is review study 13, which focused only on 6 articles.



Note: NMSSs (20) = not mentioning sample sizes; MVSSs (9) = mentioning vague sample sizes; MPSSs (3) = mentioning precise sample sizes; NMSAs (23) = not mentioning subject areas; MVSAs (5) = mentioning vague subject areas; MPSAs (5) = mentioning precise subject areas; SCD = shortest coverage duration; LCD = longest coverage duration; FAs = review study with fewer articles; MAs = review study with most articles

Figure 4 - Review studies with most and least articles, and with longest and shortest coverage duration, and studies mentioning and not mentioning subject areas and sample sizes.

There are 5 studies that mentioned precise subject areas, with natural sciences (maths and science) and computer science and engineering mentioned by 4 studies. Similarly, 5 studies mentioned vague subject areas, while 23 studies did not mention their subject areas. In this case, 3 studies provided precise sample sizes, and collectively, their sample sizes totalled 46,695 participants. Nine studies provided vague sample sizes, with 20 having not stated their sample sizes.

6.2 Purposes of the review studies

As illustrated in Figure 5, 20 review studies had purposes focusing on EDM techniques, EDM methods, EDM models, or EDM algorithms used to predict SAP and student success in higher education. Of these review studies, 15 explicitly mentioned SAP or academic/student performance in their purposes, with 3 of them mentioning both SAP and dropout prediction. Of the remaining 5 studies, 3 referred to predictive models, while 2 referred to predicting student success. The remaining 10 review studies had their purposes on reviewing or surveying EDM techniques and tools, and 2 had their purposes on student dropout prediction. The other remaining review study did not mention its purpose.

Note: EDM = educational data mining; SAP = student academic performance; SS = student success; SP = student performance

Figure 5 - Aspects identifiable from the purposes of the review studies.

6.3 Common input (predictor) variables and common output (predicted) variables employed as reported by review studies

Six typologies of input (predictor) variables emerged as the most common typologies of input variables used for predicting SAP by the reviewed studies. These are pre-university academic factors; university academic factors; student demographics; family factors; psychological factors; and student e-learning activities (Table 2). Of these collective factors, student demographics appears in 30 review studies. It is followed by both university academic factors and psychological factors. High school background and admission scores rank as the most common pre-university academic factors employed, whereas graduation percentage is reported as the commonly used aggregated attribute for university academic factors. For student demographics, gender and age are the two most common attributes reported to have been used, while family is the common attribute reported to have been employed for family factors. The common attributes for psychological factors are surveys and participation, and student discussion posts/online discussion forums are the commonly used factor for student e-learning activities.

As regards the common output variables, both university academic factors and pre-university academic factors emerged as the two frequently used attributes under these types of SAP predictor variables (Table 3).

RS 1.	To survey research trends of EDM tools, techniques & educational outcomes.
RS 2.	To provide an overview of current knowledge of LA and EDM.
RS 3.	To survey the most recent studies on EDM practices and techniques.
RS 4.	To provide an overview of DM techniques used to predict student performance; and to establish prediction algorithms that can identify the most important attributes in student data.
RS 5.	To explore EDM methods and models for improving academic performance and institutional effectiveness.
RS 6.	To survey literature in EDM in higher education and to focus on applying EMD to predict academic performance.
RS 7.	To explore the application areas and techniques of EDM, and factors affecting student academic performance.
RS 8.	To establish how performance prediction studies have evolved from those using traditional data to those utilising sophisticated data.
RS 9.	To survey different DA techniques that have been used to predict student performance and progress.
RS 10.	To determine the existing state of research on predicting student academic performance.
RS 11.	N/M
RS 12.	To present a review works in which DM techniques were used to solve educational problems and to provide a classification associated with them.
RS 13.	To offer a methodological systematic review of empirical LA research in medical education and to provide an overview of the commonly used methods.
RS 14.	To identify studies using EDM techniques to predict university dropout.
RS 15.	To provide a systematic review of university student dropout prediction through DM techniques.
RS 16.	To shed light on specific learning problems not yet addressed by previous reviews.
RS 17.	To present a comprehensive review of studies dealing with SAP and dropout predictions. **NB:** Not framed as a goal, purpose or goal).
RS 18.	To review methodological components of predictive models developed and implemented in LA applications in HE.
RS 19.	To try to comprehend a few literary works on academic performance prediction of engineering students with the focus on grade predictions.
RS 20.	To find the most critical factors affecting the student performance used by most studies; and to find the most used algorithm and the accuracy of DM algorithms.
RS 21.	To provide an overview of the current state of research activity regarding predictive analytics in HE.
RS 22.	To identify the characteristics of the MOOCs used for prediction; to describe the prediction outcomes; to classify the prediction features; to determine the techniques used to predict the variables; and to identify the metrics used to evaluate the predictive models.
RS 23.	To identify the most commonly studied factors that affect the students' performance and the most common DM techniques applied to identify these factors.
RS 24.	To identify the predictive methods for students' academic performance in HE.
RS 25	To review the latest trends in predicting students' performance in higher education.
RS 26.	To provide guidelines for educators willing to apply DM techniques to predict student success.
RS 27.	To conduct a comprehensive review of EDM studies in Turkey.
RS 28	To identify and present research published over the last five years (2015-2019) in relation to assessing students' academic performance using data mining techniques.
RS 29	To investigate explainable models of student performance prediction from 2015 to 2020.
RS 30	To find the most used algorithm by researchers in the field of supervised machine learning in the period of 2010-2020.
RS 31	To provide an overview of the current state of research in EDM.
RS 32	To obtain the most effective EDM approaches used to identify students that may underperform in computer programming.
RS 33	To create a comprehensive understanding of the landscape of academic performance prediction by focusing on the attainment of learning outcomes.

Table 1 - Purposes of individual review studies.

Pre-university academic factors	High school final grade (11); high school background (2, 4); distance high school, entrance exam (9, 32); pre-course performance, school performance (10); high school type, high school department, high school grade, admission score (17, 29, 32); linguistic features extracted from college admission application essays (18, 19, 25); teaching medium, class size, school reputation (20, (21); CGPA (23); GPA, assessment (26) (n = 16)
University academic factors	Internal assessments and external assessments, CGPA (4); end-of-semester exam, GPA, assignment, attendance, unit test, graduation, graduation percentage (5, 25, 28, 29, 32, 33); pass/fail, exact score (8); mid-term marks, lab test grade, scholarship (9, 10); drop out or not (11); behaviour in certain courses (12, 15); students' self-assessment, task complexity evaluation (16, 17); enrolment (18, 19); sessional marks (20); notes (24); achievement scores (26) (n = 18)
Student demographics	Student demographics (4, 10, 21, 28, 32); gender (5, 9, 10, 11, 17, 19, 20, 24, 25, 26); age (10, 17, 20, 23, 24, 25, 26); race, marital status, nationality (17, 18, 22, 25, 27); language, origin, educational background (24, 26, 32) (n = 30)
Family factors	Family background, parents' education (5) or father's education, father's occupation, mother's education and mother's occupation (11, 25); family (9, 10); (19); support (20); number of siblings, student's place of residence (23); (26) (n = 9)
Psychological factors	Psychometric factors (4); self-confidence, interest, course and degree ambition, participation (9, 32, 33); engagement, personality, task time, motivation, self-regulation (10); learning strategies survey, LMS questionnaires (13); (15); student preferences, planning strategies, satisfaction (16); stress management, first generation learner, learning style (17); attitude and socio-emotional surveys, teaching quality and style (18); weight (20); student effort, classroom characteristics (21); (22); instructor's knowledge and clarity, course evaluation surveys, students' environment (23); learning time (24); study behavior (25, 32); scales (26) (n =18)
Student e-learning activities	Discussion posts/online discussion forums (2, 3, 33); log data (10); students' LMS data usage, students' access data to and time usage (13, 29); student activity data from LMSs (18, 28); platform use (22); message chat logs, frequency of course clicks (23); (25); navigation data (26) (n = 12)

Table 2 - Common (input) predictor variables employed as reported by review studies.

Pre-university academic factors	Admission exam grade (6); academic background, pre-post enrolment factors (7) (n = 2)
University academic factors	Course grade, GPA, pass/fail course, semester, year, drop out or not, scholarship (6, 30, 31, 33); CGPA, GPA, class attendance, sessional marks, final grade, course content (7); course grade/score, exam/post-test grade, course grade range, pass/fail, programme/module graduation/retention, SGPA, assignment performance (e.g., grade, time to completion), course retention/dropout, knowledge gain, number of courses passed or failed (10, 30, 31, 33); risk of failing a course, dropout risk, grade prediction and graduation rate (21, 30, 33); scores prediction (22) (n = 13)

Table 3 - Common output variables employed as reported by review studies.

6.4 Common EDM methods employed as reported by review studies

There are seven commonly used EDM methods for predicting SAP as reported by the reviewed studies (Table 4). Of these, the most commonly used EMD method is classification, which is reported by 16 review studies. It is followed by clustering, which is reported by 14 review studies. Both regression and association rules are ranked third and fourth, respectively. Naïve Bayes is the least commonly used as it is referenced by only 7 review studies.

Classification	1, 2, 3, 4, 5, 6, 7, 9, 11, 12, 16, 21, 23, 24, 25 & 26 (n = 16)
Clustering	1, 2, 3, 5, 6, 7, 9, 12, 16, 23, 24, 25, 26 & 27 (n = 14)
Regression	2, 4, 5, 6, 7, 9, 16, 21, 22, 24, 25 & 26 (n = 12)
Association rule(s)	1, 2, 3, 5, 6, 7, 9, 12, 15, 16 & 26 (n = 11)
Decision Tree(s) (DT(s)	12, 14, 15, 17, 18, 19, 20, 22 & 25 (n = 9)
Support Vector Machine (SVM)	14, 15, 17, 18, 19, 20, 22 & 25 (n = 8)
Naïve Bayes (NB)	15, 17, 18, 19, 20, 22, 25 (n = 7)

Table 4 - Common EDM methods employed as reported by review studies.

6.5 Common EDM algorithms (classifiers) and common EDM software tools employed as reported by review studies

Pertaining to the commonly used EDM algorithms for predicting SAP, there are seven algorithms referenced by the reviewed studies (Table 5). Of these seven EDM algorithms, DT is the most commonly used algorithm as it is mentioned and cited by 24 review studies. It is followed by SVM (n = 20), ANN (n = 19), NB (n = 15), and K-NN (n = 13), respectively. Naïve Bayes classifiers is the least commonly used EMD algorithm for predicting SAP. However, when Bayesian classifiers are clustered together, they emerge as the most frequently utilised EDM algorithms as reported by 28 review studies.

DT (Decision Tree)	3, 4, 5, 7, 8, 9, 10, 11, 12, 14, 15, 17, 18, 19, 20, 22, 23, 25 ,27, 29, 30, 31, 32 & 33 (n = 24)
SVM	4, 5, 7, 8, 9, 14, 15, 17, 18, 19, 21, 22, 23, 24, 27, 28, 29, 30, 31 & 33 (n = 20)
ANN (Artificial neural networks)	4, 5, 7, 8, 9, 11, 14, 15, 17, 18, 19, 21, 22, 23, 27, 28, 30, 32 & 33 (n = 19)
NB (Naïve Bayes)	3, 4, 7, 9, 10, 14, 15, 17, 18, 19, 21, 22, 29, 31 & 32 (n = 15)
K-NN (K-Nearest Neighbour)	4, 5, 7, 9, 15, 17, 18, 21, 23, 24, 26, 30 & 31 (n= 13)
BN (Bayesian network)	5, 8, 11, 14, 26, 28, 31 & 33 (n = 8)
Naïve Bayes classifiers	8, 23, 24, 27 & 32 (n = 5)
Algorithm(s) reported to have the highest prediction rate	DT (n = 7)

Table 5 - Common EDM algorithms (classifiers) employed as reported by review studies.

Seven of the review studies reported on and mentioned the EDM techniques or algorithms with the highest student performance prediction accuracy rate. Of these studies, DT is reported to have the highest prediction accuracy rate by four studies (a 100% and a 99% predication accuracy rate by one study). It is followed by Naïve Bayes, which has a mixed prediction accuracy rate: two studies rate it as having a high prediction accuracy rate, one of which rates it to have a prediction accuracy rate of 100%), whereas two studies rate it as having a low prediction accuracy rate (a 76% prediction accuracy rate in one study).

In this context, three EDM software tools are reported to be frequently used for predicting SAP. These are WEKA, SPSS and RapidMiner, with WEKA as the most commonly used of the three EDM software tools (Table 6).

WEKA	1, 3, 6, 14, 15, 20, 25, 27 & 28 (n = 9)
SPSS	5, 6, 14, 15, 25 & 27 (n = 6)
RapidMiner	4, 14, 20, 25 & 27 (n = 5)

Table 6 - Common EDM software tools as reported by review studies.

7. Discussion

In this section, the discussion of the findings is structured in response to the six research questions of the study. As pointed out above, 33 review studies constituted the focal point of the present overview. Except for four studies, the rest (n = 29) were reviews of different typologies: systematic reviews (n = 9); classical reviews (n =8 systematic reviews (n = 5); surveys (n = 4); and literature reviews (n =3). In their review of reviews, Kim et al. (2018) investigated qualitative reviews (narrative and thematic reviews) and quantitative reviews (systematic and meta-analysis reviews) as part of the articles (n = 171) included in their study on hospitality and tourism.

Concerning subject areas, maths and science, and computer science and engineering featured among the subject areas mentioned by 5 studies. In this case, 3 studies mentioned sample sizes that together totalled 46,695 participants. A review of reviews in a different but related area that offers subject areas on which its reviews focused is Kim et al. (2018). Of the 13 reviews this overview reviewed, economics and finance (n = 29), customer behaviour (n = 24), and marketing (n =22) are reported as the top three subject areas mentioned by the reviewed studies, respectively. The overview mentions that sample sizes of its 171 reviews ranged from less than 10 to more than 10,000, with systematic reviews having the highest sample sizes. In the current overview, the 3 reviews that mentioned specific sample sizes were a systematic literature, a literature review, and a meta-analysis (see Figure 4 and Appendix A).

Pertaining to the purposes of the 33 reviews, it emerged that the purposes of 20 reviews had to do with either EDM techniques, EDM methods, EDM models, or EDM algorithms utilised to predict SAP and student success in higher education. By contrast, of the remaining 13 studies, 10 reviewed or surveyed EDM techniques and tools, whereas 2 focused on student dropout prediction. A study that had purposes (or objectives) as one of its focal points of analysis is Khanna et al.'s (2016) systematic review, which had reviewed 13 articles. Among the purposes of the 13 articles it analysed, educational data mining (EDM) methods or techniques employed for predicting student performance featured prominently in the purposes of 10 of these articles. The other study, Papamitsiou and Economides' (2014) systematic literature review of 40 articles, had six purposes, of which prediction of student performance was the second most common purpose after student behaviour modelling.

Of the six typologies of input variables reported to have been used by the 33 review studies, student demographics emerged as the most commonly used input variable for predicting SAP, with both gender and age as the most common attributes. It was followed by both university academic factors and psychological factors, with graduation rate, and both surveys and participation as the most common attributes for each of these two collective factors, respectively. In Khasanah's (2018) review of 10 articles, student personal information and family information were the two most popular collective factors used, with gender and age, and father education and mother education, as their most common attributes, in each case. Pre-university (high school results) and university (GPA and assessment grades) factors and student demographics (gender and age) are the most influential factors reported in Alyahyan and Düştegör's (2020) literature review of 19 articles. For output variables, both pre-university academic factors and university academic factors were the two frequently employed cluster of factors with reference to these types of SAP predictor variables.

As characterised in the findings section, the four most commonly used methods were classification, clustering, regression, and association rules, respectively, while Naïve Bayes was the least utilised method. Similarly, both classification and clustering were the most popularly used EDM methods in Papamitsiou and Economides' (2014) systematic literature review, while regression was the third most used method. Classification was found to have been the most popularly used EDM method in Ganesh and Christy's (2015) survey of 10 articles, with association rules and clustering as the second and third most used methods, successively. Again, classification was found to be the top-most utilised EDM method (n = 40) by Del Río and Insuasti's (2016) review study of 56 articles.

In relation to the seven EDM algorithms identified from the 33 review studies, Decision Tree (DT) was found to be the most commonly employed for predicting SAP, with Support Vector Machine (SVM) and Artificial Neural Network (ANN) being the second and third most used algorithms, respectively, while Naïve Bayes (NB) was the least used algorithm. Nonetheless, as a cluster, Bayesian classifiers were the most frequently utilised, overall. One review study that found DT to be the most used EDM algorithm is Cui et al.'s (2019) review of 121 articles. It was referenced by 46 of these articles, followed by Naïve Bayes (n= 32), SVM (n = 26), and neural networks (NN) and multi-layer perceptron (MLP) (n = 26). Similarly, DT had a frequency of 49 as opposed to two of its nearest algorithms, Bayesian classifiers (f = 36) and NN (f = 29) in Agrusti et al.'s (2019) systematic review of 73 studies.

In another scenario, DT and Naïve Bayesian classifiers (as categories) had the frequencies of 35 (24.8%) and 14 (9.9%) out of the total number of 141 algorithms identified from 34 articles in Saa et al.'s (2019) systematic review. However, when viewed as individual algorithms, Naïve Bayesian classifiers had the frequency of 13 (38.2%), followed by SVM with the frequency of 8 (23.5%). DT had the frequency of 4 (11.8%). In terms of the student performance prediction accuracy, only nine review studies stated EDM techniques or algorithms that had such a prediction accuracy. DT emerged as the EDM algorithm that had the highest student performance prediction accuracy rate as mentioned by 6 of the 9 studies, while Naïve Bayes had mixed prediction accuracy rates. In Ganesh and Christy's (2015) survey, DT generated the most consistent prediction results as opposed to Naïve Bayes, J48 and JRip.

Lastly, pertaining to EDM software tools for predicting SAP, WEKA emerged as the most commonly employed tool, followed by both SPSS and RapidMiner. WEKA was similarly found to be an EDM software tool used by 15 of the 20 papers (even though in one instance it was used in tandem with RapidMiner), while both RapidMiner and Matlab were each used by 3 papers in Kumar et al.'s (2018) review. In the same breath, WEKA appeared in 14 articles, followed by SPSS (n =9) and R (n = 8) and RapidMiner (n = 5) in Agrusti et al.'s (2019) in systematic review of 73 articles.

8. Conclusions, limitations and further research

The purpose of this overview was to compare and synthesise the findings of review studies conducted on predicting SAP in higher education using EDM methods, algorithms, and tools from 2013 to September 2021. For subject areas, maths and science, and computer science and engineering were cited by the review studies that explicitly mentioned their fields of study. Humanities and social sciences subjects did not feature in any of these review studies. Concerning sample size, only 3 studies explicitly stated their precise sample sizes, of which the total number was 46,695.

Among the EDM methods used for predicting SAP, four emerged as the most commonly used: classification, clustering, regression, and association rules. Classification was the most commonly used of the four methods. Naïve Bayes was the least utilised method. Of the seven commonly used EDM algorithms identified by the 33 review studies for predicting SAP, DT was the most commonly employed, followed by both Support Vector Machine (SVM) and Artificial Neural Network (ANN) respectively, with Naïve Bayes (NB) as the least used algorithm. Nevertheless, as a cluster of algorithms, Bayesian classifiers were the predominantly used algorithms. Moreover, DT was an EDM algorithm that was reported as having the highest prediction accuracy rate for predicting SAP. With respect to EDM software tools, WEKA was the most commonly utilised tool, followed by both SPSS and RapidMiner.

Finally, it is critical that future reviews on predicting SAP using EDM methods, algorithms, and tools should avoid the pitfalls identified above and those highlighted elsewhere in this overview. Most importantly, more overview studies are needed to build on the current overview study with a view to comparing and synthesising the different aspects of existing and future review studies focusing on predicting SAP using EDM methods, algorithms, and tools.

References

* = Review articles

*Agrusti, F., Bonavolontà, G., & Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. *Journal of E-Learning and Knowledge Society*, 15(3), 161–182. https://doi.org/10.20368/1971-8829/1135017

*Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: A systematic review. *IEEE Access*. 9, 33132–33143. 10.1109/ACCESS.2021.3061368

*Alban, M., & Mauricio, D. (2019). Predicting university dropout through data mining: A systematic literature. *Indian Journal of Science and Technology*, 12(4), 1–12. https://doi.org/10.17485/ijst/2019/v12i4/139729

*Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, 37, 13–49. https://doi.org/10.1016/j.tele.2019.01.007

*Alturki, S., Hulpus, I., & Stuckenschmidt, H. (2020). Predicting academic outcomes: A survey from 2007 till 2018. *Technology, Knowledge and Learning*, 1–33. https://doi.org/10.1007/s10758-020-09476-0

*Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17(3), 1–23. https://doi.org/10.1186/s41239-020-0177-7

*Anoopkumar. M., & Rahman, A. M. J. M. Z. (2016). A review on data mining techniques and factors used in educational data mining to predict student amelioration. *International Conference on Data Mining and Advanced Computing* (SAPIENCE), 122–133. https://doi.org/10.1109/SAPIENCE.2016.7684113.

*Ameen, A. O., Alarape, M. A., & Adewole, K. S. (2019). Students' academic performance and dropout predictions: A review. *Malaysian Journal of Computing*, 4(2), 278–303.

*Ashenafi, M. M. (2017). A comparative analysis of selected studies in student performance prediction. *International Journal of Data Mining & Knowledge Management Process*, 7(4), 17–32. https://doi.org/10.5121/ijdkp.2017.7402

*Aydogdu, S. (2020). Educational data mining studies in Turkey: A systematic review. *Turkish Online Journal of Distance Education*, 21(3), 170–185.

*Cui, Y., Chen, F., Shiri, A., & Fan, Y. (2019). Predictive analytic models of student success in higher education: A review of methodology. *Information and Learning Sciences*, 120(3/4), 208–227. http://dx.doi.org/10.1108/ILS-10-2018-0104

*Del Río, C. A., & Insuasti, J. A. P. (2016). Predicting academic performance in traditional environments at higher-education institutions: A review. *Ecos de la Academia*, 4, 185–201.

*Durga, V. S., & Thangakumar, J. (2019). Students performance prediction through educational data mining - An uncomplicated review. *International Research Journal of Engineering and Technology*, 1404–1406.

*Ganesh, S. H., & Christy, A. J. (2015). Applications of educational data mining: A survey. IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems (pp. 1–6). https://doi.org/10.1109/ICIIECS.2015.7192945.

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, 26, 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

*Hamoud, A. K., Dahr, J. M., Najim, I. A, & Kamel, M. B. M. (2021). Supervised learning algorithms in educational data mining: A systematic review. *Southeast Europe Journal of Soft Computing*, 10(1), 55–70.

Hamoud, A., Humadi, A., Awadh, W. A., & Hashim, A. S. (2017). Students' success prediction based on Bayes algorithms. *International Journal of Computer* Applications, 178(7), 6–12.

*Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., … Liao, S. N. (2018). Predicting academic performance: A systematic literature review. Proceedings *Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education* (pp. 175–199). ACM: Cyprus. https://doi.org/10.1145/3293881.3295783

*Jindal, R., & Borah, M. D. (2013). A Survey on educational data mining and research trends. *International Journal of Database Management Systems*, 5(3), 53–73. https://doi.org/10.5121/ijdms.2013.5304

*Khanna, L., Singh, S. N., & Alam, M. (2016). Educational data mining and its role in determining factors affecting students academic performance: A systematic review. Proceedings of India International Conference on Information Processing (IICIP) (pp. 1–7). https://doi.org/10.1109/IICIP.2016.7975354.

*Khasanah, A. U. (2018). (2018). A review of student's performance prediction using educational data mining techniques. *Journal of Engineering and Applied Sciences*, 13(6), 5302–5307. https://doi.org/10.36478/jeasci.2018.5302.5307

Kim, C. S., Bai, B. H., Kim, P. B., & Chon, K. (218). Review of reviews: A systematic analysis of review papers in the hospitality and tourism literature. *International Journal of Hospitality Management,* 70, 49–58. https://doi.org/10.1016/j.ijhm.2017.10.023

Kumar, A. D., Selvam, R. P., Kumar, K. S. (2018). Review on prediction algorithms in educational data mining. *International Journal of Pure and Applied Mathematics,* 118(8), 531–537.

Kumar, M., & Salal, Y. K. (2019). Systematic review of predicting student's performance in academics. *International Journal of Engineering and Advanced Technology,* 8(3), 54–61.

*Kumar, M., Singh, A. J., & Handa, D. (2017). Literature survey on student's performance prediction in education using data mining techniques. *International Journal of Education and Management Engineering,* 6, 40-49. https://doi.org/10.5815/ijeme.2017.06.05.

*Liz-Domínguez, M., Caeiro-Rodríguez, M., Llamas-Nistal, M., & Mikic-Fonte, F. (2019). Predictors and early warning systems in higher education - A systematic literature review. Proceedings of *LASI Spain 2019: Learning Analytics in Higher Education* (pp. 84–99). http://ceur-ws.org/Vol-2415/paper08.pdf.

*López-Zambrano, J., Torralbo, J. A. L.; & Romero, C. (2021). Early prediction of student learning performance through data mining: A systematic review. *Psicothema*, 33(3), 456–465.

*Manjarres, A. V., Sandoval, L. G. M., & Suárez, M. J. (2018). Data mining techniques applied in educational environments: Literature review. *Digital Education Review,* 33, 235–266. https://doi.org/10.1007/s10489-012-0374-8

Miles, M. B., & Huberman, A M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, Sage.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *British Medical Journal,* 339:b2535, 1–8. https://doi.org/10.1136/bmj.b2535

*Moonsamy, D., Naicker, N., Adeliyi, T. T., & Ogunsakin, R. E. (2021). A meta-analysis of educational data mining for predicting students performance in programming. *International Journal of Advanced Computer Science and Applications*, 12(2), 97-104.

*Moreno-Marcos, P. M., Alario-Hoyos, C., Munoz-Merino, P. J., & Kloos, C. D. (2019). Prediction in MOOCs: A review and future research directions. *IEEE Transactions 0n Learning Technologies,* 12(3), 384–401. https://doi.org/10.1109/TLT.2018.2856808

*Muttathil, A., & Rahman, A. M. J. (2016). A review on data mining techniques and factors used in educational data mining to predict student amelioration. 2016 International Conference on Data Mining and Advanced Computing. https://doi.org/10.1109/SAPIENCE.2016.7684113

*Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11, 237.

*Papadogiannis, I., Poulopoulos, V., & Wallace, M. (2020). A critical review of data mining for education: What has been done, what has been learnt and what remains to be seen. *International Journal of Educational Research Review*, 5(4), 353–372.

*Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society,* 17(4), 49–64.

Pieper, D., Buechter, R., Jerinic, P., & Eikermann, M. (2012). Overviews of reviews often have limited rigor: A systematic review. *Journal of Clinical Epidemiology*, 65, 1267–1273. http://dx.doi.org/10.1016/j.jclinepi.2012.06.015

Polanin, J. R., Maynard, B. R., & Dell, N. A. (2016). Overviews in education research: A systematic review and analysis *Review of Educational Research*, 87(1), 172–203. https://doi.org/10.3102/0034654316631117

*Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, 24, 567–598. https://doi.org/10.1007/s10758-019-09408-7

*Saqr, M. (2018). A literature review of empirical research on learning analytics in medical education.

International Journal of Health Sciences, 12(2), 80–85.

*Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Vaismoradi, M., & Snelgrove, S. (2019). Theme in qualitative content analysis and thematic analysis. *Forum: Qualitative Social Research*, 20(3), Art. 23.

*Zulkifli, F., Mohamed, Z, & Azmee, N. A. (2019). Systematic research on predictive models on students' academic performance in higher education. *International Journal of Recent Technology and Engineering*, 8(2S3), 357–363. https://doi.org/10.35940/ijrte.B1061.0782S319