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

Appendix A. Summary of key aspects of each review article

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
1. Jindal & Borah (2013)	India	Survey	1998-2012	37 articles	N/M	N/M	To survey research trends of EDM tools, techniques & educational outcomes.	Web data (1998-2000); institutional data ((2001-2004); survey data (2005-20008); public repository data (2009-2012)	N/M	Web mining (1998-200); association mining (2001-2004); classification-DT, clustering & association mining (2005-2008); SVM & neural network (2009-2012). **NB:** Relationship (52%); prediction (28%); exploratory data analysis (17%); cluster analysis (15%)	DB Miner; WebSIFT; MS SQL Server; Oracle Data Miner; WEKA; SPSS Clementine	N/M	Different MD techniques were in common use during the different time spans, with classification-DT, clustering & association mining as the most used techniques. WEKA and SPSS Clementine were the most preferred tools between 1998-2012.
2. Papamitsiou & Economides (2014)	Greece	Systematic literature review	2008-2013	40 studies	VLEs/ LMSs, MOOCs, cognitive tutors, multimodality & mobility	N/M	To provide an overview of current knowledge of LA and EDM.	Log files; chat messages; response times; resources accessed; previous & final grades; discussion posts, student profiles; Google analytics; open datasets; virtual machines	N/M	Classification (20); clustering (7); regression (3); discovery with models (3); visualisation (3); text mining (3); association rule (2); SNA (2); statistics (2)	Classification (20); clustering (7); regression (3); discovery with models (3); visualisation (3); text mining (3); association rule (2); SNA (2); statistics (2)	N/M	Unrelated to EDM techniques
3. Ganesh & Christy (2015)	India	Survey	2009-2014	10 articles	N/M	N/M	To survey the most recent studies on EDM practices and techniques.	Students' performance prediction; student performance via online discussion forums; student dropout rate; student retention rate; teacher's class questions; online education video behaviour; student profile; e-learning system activities;	N/M	Classification (5); association rules (3); clustering (2); visualisation (2); feature selection (1)	Naïve Bayes (2); J48 (2); a priori algorithm (2); random tree (1); JRip (1); EM (1); feature selection (e.g., term frequency, mutual information, information gain & Chi Square) (1); classification (e.g., K-NN, Naïve Bayes, SVM & Rochio	For classification, DT produced consistent results (100% in two datasets & 99% for one dataset) for prediction accuracy compared	EDM contributes to improving HE. DT generated consistent results for classification while J48, JRip & Naïve Bayes produced inconsistent results.

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
								learning through social networking distance learning final performance. **NB:** Each factor relates to each article.			Algorithm) (1); Grapviz (1); K-Means (1); WEKA (1), Genetic Algorithm (1)	with J48, JRip & Naïve Bayes.	
4. Shahiri et al. (2015)	Malaysia	Review	2002-early 2015	30 papers	N/M	University students (**NB:** Number not mentioned)	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.	Internal assessments; external assessments; psychometric factors; CGPA; student demographics; high school background; scholarship; social network interaction; & extra-curricular activities	N/M	Classification (N/M), regression (N/M) & categorisation (N/M)	DT (10); ANN (8); Naïve Bayes (4); K-NN (3) & SVM (3)	ANN had the highest prediction accuracy (98%) and is followed by DT (91%). Naïve Bayes had the lowest prediction accuracy (76%).	The most used variables/datasets were CGPA and internal assessment; classification was the most frequently used EDM method; and ANN and DT were the two most common algorithms with the former having the highest prediction accuracy for student performance.
5. Anoopkumar & Rahman (2016)	India	Review	2005-2015	40 papers	N/M	N/M	To explore EDM methods and models for improving academic performance and institutional effectiveness.	Some of the factors mentioned are: gender; family background; parents' education; end-of-semester exam; GPA; CGPA; assignment; attendance; unit test; graduation percentage, etc.	N/M	EDM (4); classification (23); clustering (6); association (6); sequential mining (1); text mining (1); interactive mining (1); temporal mining (1); ANN (1); distributed DM (1); web mining (1); regression (3); correlation (3); statistical analysis (10); visualisation (10)	Bayesian Network; DT; ANN; SVM; K-NN	N/M	Student academic performance (SAP) prediction featured in 20 papers, while 20 papers focused on EDM techniques.

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
6. Del Río & Insuasti (2016)	Ecuador	Review	2011-August 2016	56 + 5 articles	N/M	N/M	To survey literature in EDM in higher education and to focus on applying EMD to predict academic performanc e.	Academic and other data (29); academic data only (21); non-academic data (2); partial grades/other data (2) and partial grades only (2)	Course grade (20); some form of GPA (16); pass/fail course, semester, year (10); admission exam grade (4); job placement (2); drops out or not (1); wins scholarship (1); loss of academic status (1); student potential (1)	Classification (40); clustering (5); association rule mining (4); linear regression (3); machine learning (2) & matrix factorisation (2)	WEKA only (16); WEKA & other software (2) SPSS alone & with other software (3); SAS Enterprise Miner (1); unknown (34);	N/M	Classification was found to be the most popular method employed by the reviewed articles, followed by clustering and association. When using these methods, the need for human intervention should not be ignored. WEKA served as the software of choice.
7. Khanna et al. (2016)	India	Systematic review	2010-2015	13 publications (8 journals, 4 conferences & 1 book)	N/M	N/M	To explore the application areas and techniques of EDM, and factors affecting student academic performanc e.	N/M	CGPA (1); GPA (1); Academic background (1); family closeness)1); freedom to make choices (1); pre-post enrolment factors (1); employability (1); class attendance (1); assignment (1); sessional	Classification (4); association rule (2); regression (1); clustering (1); sequential pattern (1); relationship mining (1); prediction (3); structure discovery (1); distillation (1); discovery (1);	ANN; DT; SVM; K-NN; Naïve Bayes (1)	N/M	Classification was one of the most commonly used techniques; no generalised tools used in EDM yet/

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
									marks (1); final grade (1); course content (1)				
8. Ashenafi (2017)	Italy	Comparative analysis	Studies published since the nineties	46 studies	Computer science (30.4%); other (unreported) (30.4%); engineering (17.4%); maths/physics (8.7%); business; medicine; multiple courses; biology; psychology	Undergraduate (65.2%); other (unreported) (23.9%); both (6.5%); & graduate	To establish how performance prediction studies have evolved from those using traditional data to those utilising sophisticated data.	Grade only (37%); pass/fail only (26.1%); exact score only (17.4%); pass/fail and grade (13%); grade and exact score; pass/fail and exact score	N/M	N/M	Multiple (26.1%); ANNs (19.6%); linear regression (13%); DTs (13%); SVMs (10.9%); Random forest; Naïve Bayes classifiers; Bayesian networks; Markov networks; latent Dirichlet analysis & custom	N/M	Most commonly used algorithms: ANNs; linear regression; SVMs; Naïve Bayes classifiers; & DTs. Least used algorithms: Markov networks; collaborative multi-regression; & sentiment analysis. Much of (student) performance prediction studies have been conducted in computer science and engineering. Student demographic data and high school grades were the most common independent variables, while GPAs or CGPAs serve as dependent variables.
9. Kumar et al. (2017)	India	Survey	2007-July 2016	16 papers	Universities, schools & colleges	University, engineering institution students (**NB:**	To survey different DA techniques that have been used to predict student performanc	Miscellaneous factors and attributes: e.g., internal assessment test grade; institutional internal data sources; external data sources; assignment	N/M	Classification; clustering; association rules; regression	DT; Naïve Bayes; SVM; ANN; K-NN; rule-based algorithms; K-NN; Random forest; Random tree; SMO; REPTree; LADTree, J48	DT, Naïve Bayes and K-NN were found to have the highest prediction accuracy (100%)	CGPA and internal marks were important attributes for predicting student academic performance. Most studies employed DT, Naïve Bayes

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
						Number not mentioned)	e and progress.	submission and grade; correct response; self-confidence; interest; course and degree ambition; mid-term marks; lab test grade; attendance; participation; gender; family; distance high school; CGPA; GPA; entrance exam; scholarship; etc.				followed by rule-based algorithms. ANN was found to have the lowest prediction accuracy (89.8%).	and rules-based algorithms for predicting student academic performance.
10. Hellas et al. (2018)	Multiple countries (Finland, Canada, Macedonia, Australia & USA)	Systematic literature review	2010-2017	357 articles	Computer science (126/34.9%); STEM (98/27.1%); other (39/10.8%); multidisciplinary (30/8.3%); unclear (14/3.9%)	Post-secondary students	To determine the existing state of research on predicting student academic performance.	Miscellaneous attributes: course performance (141/13.09%); pre-course performance (139/12.91%); engagement (113/10.4%); gender (86/7.99%); personality (65/6.04%); demographic (65/6.04%); school performance (58/5.39%); age (53/4.92%); family (52/4.83%); task time (41/3.81%); motivation (33/3.06%); self-regulation (28 (2.60%); log data (28/2.60%); etc.	Miscellaneous values: course grade/score (88/24.4%); exam / post-test grade or grade (53/14.7%); course grade range (e.g., A-B, Pass/Fail) (49/13.6%); programme / module graduation / retention (48/13.4%); vague / unspecificied performance (44/12.2%); GPA or GPA range	Statistical linear modelling (110/17.71%); probabilistic graphical model (80/12.88%); classification: DTs (74/11.92%); statistical: correlation (57/9.18%); classification: NN (51/8.21%); classification: SVM (45/7.25%); classification: classification (42/6.76%); statistical: latent variable models (27/4.35%); classification: random forest (25/4.03%); clustering: Partition-based (19/3.06%); classification:	See the preceding column.	N/M	The majority of articles reviewed (38%) used individual course grade as prediction metric, while 11.4% of the articles focused on assignment performance. The mostly used EDM methods were classification (e.g., Naïve Bayes and DTs) and clustering (e.g., partitioning data), statistical analysis (e.g., correlation and regression), and data mining.

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
									(plus CGPA, SGPA) (44/12.2%); assignment performanc e (e.g., grade, time to completion) (41/11.4%); course retention / dropout (20/5/5%); knowledge gain (8/2.2%); number of courses passed or failed (4/1.1%)	nearest neighbour (17/2.74%); etc.			
11. Khasanah (2018)	Indonesia	Review	2007-2010	10 articles	N/M	N/M	N/M	Personal data (e.g., gender, origin); family data (e.g., father's education; father's occupation, mother's education, mother's occupation, high school type); pre-university data (e.g., high school department, high school final grade); university data (e.g., first semester attendance, final GPA (FGPA), drop out or not)	N/M	classification	DT (8); Bayesian network (5); NN (1); other (1)	N/M	DT and Bayesian network emerged as the most used methods for predicting student performance. DT outperformed the other methods with thee CART algorithm. Most widely used attributes for predicting student performance were: student personal data; family data; pre-university data; and university data.

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
12. Manjarres et al. (2018)	Colombia	Literature review	1993-2015	127 papers	Learning patterns identification (n=32); student patterns identification (n=31); VLE (n=29); student prediction (n=22); student performance and evaluation (n=21); educational recommendations (n=16); and student dropout or retention (n=10)	N/M	To present a review works in which DM techniques were used to solve educational problems and to provide a classification associated with them.	Factors related to: learning patterns identification (n=32); student patterns identification (n=31); VLE (n=29); student prediction (e.g., final grades, performance or behaviour in certain courses, etc.) (n=22); student performance and evaluation (n=21); educational recommendations (n=16); and student dropout or retention (n=10)	N/M	Association rules (40); clustering (29); DTs (28); sequential patterns (18); classification (17); Bayesian networks (11); NN (11)	N/M	N/M	The most commonly used DM techniques were: association rules; clustering; DTs; and sequential patterns. The domains mostly analysed were learning pattern identification; VLE; student patterns identification; student dropout.
13. Saqr (2018)	Saudi Arabia	Literature review	2016-2017	6 articles	N/M	N/M	To offer a methodological systematic review of empirical LA research in medical education and to provide an overview of the commonly used methods.	Students' LMS data usage (1); LMS data and learning strategies survey (1); students' access data to and time usage of the online anatomy cases (1); process data from online radiograph case simulation (1); LMS data and SNA (1); LMS data and questionnaires (1)	N/M	Descriptive statistics and correlation with multiple regression (1); descriptive statistics, ANOVA and correlation tests (1); descriptive statistics, pattern and time analysis, and qualitative analysis (1); descriptive statistics, visualisation, time analysis and regression (1); correlation tests, linear regression,	N/M	N/M	Mostly, the methods used were descriptive statistics, correlation tests and regression. Patterns of online behaviour and usage, and predicting achievement were the most investigated outcomes.

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
										and binary logistic regression (1); descriptive statistics (1)			
14. Agrusti et al. (2019)	Italy	Systematic review	1999-2019	73 studies	N/M	University(Not explicitly stated)	To identify studies using EDM techniques to predict university dropout.	N/M	N/M	DT (n=49); Bayesian classification (n=36); NN (n=29); logistic regression (2n=5); SVMs (17); miscellanea (n=11); K-NN (n=9)	Bayesian classification algorithms: Naïve Bayes (n=25); Bayesian network (n=7; others (n=18). NN algorithms: multilayer perception (n=11); others (n=7). SVM algorithms: Averaged perception (n=2); others (3). Logistic regression algorithms: others (3). Miscellanea algorithms: ONE R (n=4); K-means (n=3); others (n=7). DM tools: WEKA (n=14); SPSS (n=9); R (n=8); Rapid Miner (n=5); others (n=15)	N/M	The following EDM techniques were identified as having the higher use frequency: DT (67%); Bayesian classification (49%); Neural networks (40%); and logistic regression (34%). The most used DM tools were WEKA, SPSS and R.
15. Alban & Mauricio (2019)	Ecuador & Perú	Systematic literature (review)	2006-2018	67 papers	N/M	University students (Not explicitly stated)	To provide a systematic review of university student dropout prediction through DM techniques.	112 factors affecting university dropout: personal factors (n=31); academic factors (n=40); economic factors (n=15); social factors (n=21); and institutional factors (n=4).	N/M	DT (23); logistic regression (20); linear regression (18); NN classifier (14); SVM (11); Naïve Bayes (10); K-NN classifier (2); Radial basic function neighbour (2); classification association rules (1); fuzzy inference (1) rule induction (1); discriminant	See the preceding column. **NB**: EDM tools with statistical techniques: SPSS (n=6); WEKA (n=4); Matlab (n=2) **NB:** EDM tools with AI techniques: WEKA (n=26); SPPS Modeler (n=5); Matlab	Artificial techniques had greater accuracy rates.	Statistical technique had a higher frequency of use, whereas artificial techniques had greater accuracy rates. The most used DM tools were WEKA and SPSS.

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
										analysis 91); probit analysis (1)	(n=4); Rapid Miner (n=4); SAS Enterprise (n=2) others (n=4).		
16. Aldowah et al. (2019)	Malaysia & Oman	Review and synthesis	2000-2017	402 studies	Four dimensions: computer-supported learning analytics (CSLA); computer-supported predictive analytics (CSPA); computer-supported behavioural analytics (CSBA); and computer-supported visualization analytics (CSVA)	Higher education students (**NB:** Number not mentioned)	To shed light on specific learning problems not yet addressed by previous reviews.	The study provides aspects such as: SNA; student preferences; students' self-assessment; task complexity evaluation; engagement; participation; planning strategies; motivation; satisfaction; etc.	N/M	Classification (26.25%), clustering (21.25%), visual data mining (155), statistics (14.25%), association rules (14%), regression (10.25%), sequential pattern mining (6.50%), text mining (4.75%), correlation mining (3%), outlier detection (2.25%), causal mining (1%) & density estimation (1%)	N/M	N/M	EDM and LA were found to be commonly used to solve learning problems. The most commonly used EDM techniques across the four dimensions were: clustering, association rule, visual data mining, statistics and regression.
17. Ameen et al. (2019)	Nigeria	Review	2007-2019	39 studies	N/M	N/M	To present a comprehensive review of studies dealing with SAP and dropout predictions. **NB:** Not framed as a goal, purpose or goal).	Personal features (e.g., age, gender, etc.); psychological features (e.g., stress management, first generation learner, learning style, etc.); academic features: pre-university academic features (e.g., high school grade, admission score, etc. & university academic features (e.g., final	N/M	Miscellaneous DM techniques and a combination of DM techniques: Naïve Bayes (n=19); SVM (13; DT (n=9); J48 (8); K-NN (7); Neural networks (7); CART (n=6); etc.	See the preceding column.	N/M	The major concerns about SAP and dropout prediction studies are related to the nature of the attributes employed in DM techniques. There is no standardisation of these techniques yet.

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
								grade, GPA, CGPA, course marks, etc.); social features (e.g., number of friends, sporting, extra-curricular activities, etc.); economic features (e.g., family income, parent's education, financial aid from third parties, etc.; demographic features (e.g., marital status, race, nationality, etc.)c					
18. Cui et al. (2019)	Canada & China	Review	2002-early 2018	121 articles	N/M	N/M	To review methodological components of predictive models developed and implemented in LA applications in HE.	Course-level prediction: course & mid-term marks; student activity data from LMSs; attitude and socio-emotional surveys and questionnaires; demographics & previous academic history; course, modality, discipline & enrolment; teaching quality and style; programme-level prediction: demographics & previous academic history; Facebook & Twitter data; linguistic features extracted from college admission application essays.	N/M	DT (n=46); Naïve Bayes (n=32); SVM (n=26); NN and MLP (n=26); RF (n=23); logistic regression (n=22); K-NN (n=16); other (n=25)	See the preceding column.	RF, logistic regression, Naïve Bayes classifiers tended to be good options for predictive LA applications.	The most frequently used and successful techniques were DT, Naïve Bayes classifier, SVM, ANNs, RF, and logistic regression. The most popular technique was DT (n=46).

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
19. Durga & Thangakumar (2019)	India	Survey	2013-2018	19 articles (**NB:** Not explicitly stated and counted as 20 in Table 4).	N/M	N/M	To try to comprehend a few literary works on academic performance prediction of engineering students with the focus on grade predictions.	Miscellaneous attributes: previous marks; high / secondary school grades; class test marks; class attendance; family annual income; fathers' education; mothers' education; gender; marital status; lab performance; CGPA; internal marks; external marks; etc.	N/M	Naïve Bayes (n=2); NN (n-2); SVM (n=2); DT (n=2); fuzzy (n=2); optimisation techniques (n=2)	See the preceding column.	DT had the highest prediction accuracy in 4 articles.	The reviewed studies employed miscellaneous factors to predict academic performance and student grades.
20. Kumar & Salal (2019)	India & Russia	Systematic review	2012-2017	58 articles (**NB:** Not explicitly stated and counted as 20 in Table 4).	N/M	N/M	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.	Miscellaneous attributes: such as academic attributes (e.g., internal and external assessment, lab marks, sessional marks, attendance, CGPA, semester marks, grade, school marks, etc.); personal attributes (e.g., age, gender, student interest, weight, level of motivation, etc.); family attributes (e.g., qualification, occupation, income, support, siblings, etc.); social attributes (e.g., number of friends, social network, movies, etc.); school attributes (e.g., teaching medium,	N/M	DT; NN; Naïve Bayes; K-NN; & SVM.	WEKA; RapidMiner; MATLAB; KNIME; Rattle GUI; Orange; Apache Mahout; R; ML-Flex; NLP Toolkit; etc.	DT had the highest prediction accuracy followed by NN and SVM. Naïve Bayes had the lowest prediction accuracy.	CGPA and internal and external assessment marks were the attributes used most by the reviewed articles. Classification, clustering, linear regression and association rules DM methods used, with classification as the most used method. WEKA was the most used DM prediction tool followed by RapidMiner.

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
								class size, school reputation, etc.)					
21. Liz-Domínguez et al. (2019)	Spain	Systematic literature review	2012-2019	26 documents / applications	Various HE environments	31,274[1]	To provide an overview of the current state of research activity regarding predictive analytics in HE.	Student demographics & background (n=6); student engagement & effort (n=15); performance & academic history (19); course, degree, or classroom characteristics (n=4); others (n=6).	Risk of failing a course; dropout risk; grade prediction; and graduation rate	Classification (n=18); regression (n=8)	Naïve Bayes; logistic regression; RF; K-NN; SVM; & NN	N/M	The most commonly used classifiers were Naïve Bayes; logistic regression; RF; K-NN; SVM; & NN. The selected predictors had a diversity in terms of their contexts, input data, prediction algorithms and prediction goals.
22. Moreno-Marcos et al. (2019)	Spain	Review	2014-2017	88 papers	Professions & applied sciences (n=46); social sciences (n=31); formal sciences (n=27); humanities (n=17); & natural sciences (n=14)	N/M	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	Demographics (n=17); video-related features (n=42); exercise-related features (n=45); forum-related features (n=46); platform use (n=52); `survey (n=8); others (n=14)	Dropout (n=34); scores prediction (n=15); certificate earners (n=14); student behaviour (n=14); relevance of content (n=5); others (n=5)	Regression (n=47); SVM (n=27); RF (n=18); DTs (=14); NNs (n=14); gradient boosting (n=11); Naïve Bayes (n=7); others (n=42)	See the preceding column.	N/M	There is strong interest in predicting dropouts in MOOCs. A variety of predictive models are used, though regression and SVM stand out. There is also wide variety in the choice of prediction features, but clickstream data about platform use stands out.

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
							predictive models.						
23. Saa et al. (2019)	UAE & Vietnam	Systematic review	2009-2018	34 research articles	N/M	N/M	To identify the most commonly studied factors that affect the students' performance and the most common DM techniques applied to identify these factors.	Students' previous grades & class performance (e.g., high school marks, CGPA, etc.) (26%); students' e-learning activities (e.g., message chat logs, system logs for a virtual room, etc.) (25%); students' demographics (e.g., age, number of siblings, student's place of residence, etc.) (23%); students' social data (e.g., smoking habits, studying groups, etc.) (12%); instructor attributes (e.g., instructor' knowledge, clarity, etc.) (4%); course attributes (3%); course evaluations (e.g., frequency of course clicks, course evaluation surveys, etc.) (3%); students' environment (2%)	N/M	Classification (n=34); clustering (n=4)	Naïve Bayes classifiers (n=13/38.3%); SVM (n=8/23.5%); logistic regression (n=17.6%); K-NN (n=5/14.7%); ID3 Decision tree (n=4/11.8%); C4.5 Decision tree (n=4/11.8%); DT (n=4/11.8%); MLP neural network (n=4/11.8%); NN (n=4/11.8%)		The most widely used factors for predicting student performance in HE are: students' previous grades and class performance, students' e-learning activities, students' demographics, and students' social data. The most common DM techniques used to predict and classify students' factors are DTs, Naïve Bayes classifiers, and ANNs.
24. Zulkifli et al. (2019)	Malaysia	Systematic literature review	2014-2018	69 articles	N/M	N/M	To identify the predictive methods for students' academic performance in HE.	Academic factors (e.g., attendance, learning time, learning activities, notes, teaching methods, lab work, tests, assignments, etc.) (n=27); academic factors & demographics (e.g.,	N/M	Classification (n=33); regression (n=19: clustering (n=3); classification & regression (n=11; clustering & regression (n=3)	Bayes classification (n=3); K-NN (n=6); logistic regression (n=5); SVM (n=3); classification trees (n-=8); principal component analysis (n=1); regression analysis (n=14)	N/M	Predictive results using classification and cluster methods tend to predict SAP based on predetermined class, not by following the performance of

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
								gender, age, race, language, origin, educational background, etc.) (n=24); academic factors & personality factors (n=15; academic factors, demographics & personality factors (n=3).					students involved. Classification methods were the most used methods.
25. Alturki et al. (2020)	Germany	Survey	2007–2018	22 articles	NM	NM	To review the latest trends in predicting students' performanc e in higher education.	Gender (n = 14); GPA (n = 12); course grades (n = 10); age (n = 9); language proficiency (n = 7); income (n = 6); nationality (n = 4); marital status (n = 4); employment status (n = 4); & attendance (n = 3)	NM	DT (n = 18); Bayesian-tree (n = 5); SVM (n = 2); K-NN (n = 3); NB (n =7); Random Forest (n = 1); rule induction (n = 1); bagging (n =1); clustering (n =1); & logistic regression (n = 2)	Weka (70%), SPSS (15%), RapidMiner (10%); & others (5%)	DT algorithms (especially C4.5) reported to have the highest accuracy rate.	DT methods (C4.5, CART, ADT, CHAID and ID3) were the most used algorithmic methods during the period under review. Weka was reported to be the most used tool. It was followed by SPSS and RapidMiner tools. Gender, age, previous GPA and language proficiency were the most used predictor features.
26. Alyahyan & Düştegör (2020)	Saudi Arabia	Literature review	Articles published in the last 5 years	19 articles (**NB:** Not explicitly mentione d)	N/M	13,465[2]	To provide guidelines for educators willing to apply DM techniques to predict student success.	Prior academic achievement (e.g., pre-university data, high school background, GPA/CGPA, assessment grade, etc.) (44%); demographics (e.g., gender, age, race, parents' education,	N/M	Classification; regression; clustering	Classification: DT (e.g., J48, C4.5, Random tree & REPTree (44%); Bayesian algorithms (19%); ANNs (10%); rule learner's algorithms (9%); ensemble learning (75); K-NN (5%); Regression:	N/M	Prior academic achievement factors were the most used factors for predicting student academic success. Classification was the most used prediction technique with the

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
								family income, etc.) (25%); student's environment (e.g., class type, semester duration, programme type) (17%); psychological factors (e.g., student interest, study behaviour, stress, motivation, etc.) (11%); student e-learning activity (e.g., login time numbers, task numbers, test numbers, discussion board entries, etc.) (3%)			regression (3%); Clustering: X-means (2%) **NB:** Commonly used DM software tool: WEKA		highest number of algorithms. The most commonly used prediction software tool was WEKA.
27. Aydogdu (2020)	Turkey	Systematic review	No date range but the search process ended in July 2019. **NB:** The first reviewed paper was published in 2004.	48 studies (graduate theses & articles)	N/M	University students (82.6%) Secondary & high school students (17.39 %)	To conduct a comprehensive review of EDM studies in Turkey.	Achievement scores (20); surveys (12); database (10); demographics (7); navigation data (5); & scales (4)	N/M	Prediction (46.77%); classification (24.19%); clustering (19.35%); & association rules (9.68%)	ANN (21); DT (17); clustering (13); regression (8); association rules (6); BC (5); SVM (4) **NB:** Analysis tools: SPSS (8); MATLAB (5); SPSS Clementine (5); Developed in the study process (4); WEKA (4); RapidMiner (3); R programming (1); SAS Enterprise Manager (1); Others (5x1 each)	N/M	ANNs were the most used technique in most studies for predicting student achievement. Most studies aimed at predicting student achievement. Achievement scores served as data source. SPSS served as a preferred analysis tool.
28. Papadogiannis et al. (2020)	Greece	Critical review	2015-2019	120 articles	NM	NM	To identify and present research published	Student grades (33.94%); student demographics (No % given); student	NM	DT (n = 107); Bayesian methods (n = 51); ensemble learning (n = 39);	WEKA; Bayesian algorithms; Neural Networks, Support Vector Machines; &	DT algorithms had the	DT algorithms Bayesian algorithms had a usage frequency in

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
							over the last five years (2015-2019) in relation to assessing students' academic performanc e using data mining techniques.	activity data (No % given)		NN (= 33); SVM (n = 29); decision rules (n = 23); instance based (n = 19); logistic regression (n = 11); linear regression (n =11); proposed and other algorithms (n =11); & association rules (n = 5)	Ensemble Learning Methods (No % given)	highest accuracy	the articles studied. NB was used as a benchmark for comparing accuracy with the other algorithms. DT algorithms had the highest accuracy, with C4.5 having the highest accuracy of all the DT algorithms (e.g., ID3, CART, and Random Trees).
29. Alamri & Alharbi (2021)	Saudi Arabia	Systemati c review	2015-2020	15 articles	NM	NM	To investigate explainable models of student performanc e prediction from 2015 to 2020.	Mixed (n = 9); pre-course performance (n = 3); course performance (n = 2); & e-learning analytics (n = 1)	NM	Classification (n = 13) & regression (n = 2)	DT algorithms: = CART (n = 2) J48 (n = 2; Jrip (n = 1); Random Forest (n = 4) unspecified (n = 1). Rule learning algorithms: CN2 (n = 1); classification association rule (n = 1); genetic-based algorithms (n = 3). Deep learning: LSTM (n =1); SVM (n =2); NB (n = 1); Logit (n =1); & RBF (n = 1).	NM	Socio-economic features and pre-course performance were the top predictors used in the 15 studies. DT and rule based learning algorithms were the commonly used algorithms.
30. Hamoud et al. (2021)	Iraq & Germany	Systemati c review	2010-2020	90 studies	NM	NM	To find the most used algorithm by researchers in the field of supervised machine learning in	NM	Student dropout (n = 8); degrees (n = 6); student activities and background (n =6);	NM	DT; ANN; SVM; logistic regression; ZeroR; K-NN; linear classifier; ensemble model; genetic programming; conditional random fields; NN;	NM	DT algorithms were the most used EDM algorithms, and they were followed by ANN and NB algorithms. The least used algorithms were

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
							the period of 2010-2020.		student skills and performance (n = 4); course selection and completion (n = 4); learner adaptation system (n = 3); job after graduation (n= 1); student profiles (n = 1); fast response learners (n =11); future educational events (n = 1); instructor performance (n = 1) & graduation rate (n = 1)		association rules mining.		SVM, logistic regression, and K-NN.
31. López-Zambrano et al. (2021)	Spain	Systematic review	1992-Nov. 2020	82 articles	NM	Tertiary level (n = 76); secondary school level (n = 6)	To provide an overview of the current state of research in EDM.	NM	Pass/ Fail, Success/ Failure, or Retain/ Dropout (No % given)	Classification (n = 50); regression (n = 33); clustering (n = 13); association (n = 2); 7 other/not specified (n = 20)	Classification: DT (J48) (n = 31/38%); Random Forest (n = 25/30%); SVM (n = 21/26%); NB (n = 14/17%); K-NN (n = 10/12%); Boosted Trees (n = 7/(9%); Adaptive Boosting (n =7/9%); Gradient Boosting (n = 4%); & other (n = 5/6%) Regression: Logistic Regression (n =	NM	Classification was the most commonly used technique, followed by regression. The most commonly used predictive algorithms were: J48, Random Forest, SVM, and Naive Bayes (classification), and logistic and

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
											23/28%); Linear Regression (n = 12/15%); Bayesian Adaptive Regressive Tree (n = 1/1%); & other (n = 12/15%) Clustering: K-Means clustering (n = 2/2%) & Balanced Iterative Reducing and Clustering (n = 1/1%) Association: Class Association Rule (n = 1/1%) & Random Guess (n = 1/1%)		linear regression (regression). The most important factors in early prediction were student assessment and data obtained from student interaction with learning management systems.
32. Moonsamy et al. (2021)	South Africa	Meta-analysis	January 2010-November 2020.	11 articles	Introductory computer programming	1,956	To obtain the most effective EDM approaches used to identify e students that may underperform in computer programming.	Grades (e.g., in mathematics, physics, and English); entrance tests; student background factors; student demographics; student behaviour; past educational information; student programming behaviour; comfort level; language (English and Malay) proficiency; and personality factors	NM	Hybrid (n = 2); data mining (n = 8); & machine learning	PART classifier – algorithm (n =1); Multiple Back-Propagation (MBP) algorithm (n =1); Naïve Bayes (n = 3); DT (J48) (n = 4); Bayesian classifier (n =1); Multilayer Perceptron (n =1); SMO (n = 1); REPTree (n = 1); NN (n =1); CBA algorithm (n = 1); CART (n = 1); Best-First Tree (BF Tree) (n = 1); clustering and association rule (n = 1).	NM	The minimum performance of algorithm prediction was 10% and it was found in studies performed with drop out and retention. In contrast, the maximum algorithm prediction performance was found to be 36%, in a study performed with the associated student-related sub group data.
33. Namoun & Alshangiti (2021)	Saudi Arabia	Systematic literature review	2010-2020	62 articles	STEM (53.22%); not specified (NS) (26%); social	University (72.58 %); school (25.81 %); &	To create a comprehensive understanding of the landscape of	Student online learning activities, term assessment grades, and student academic emotions	Performance classes (n = 34); achievement / grade score (n =20);	NM	Statistical models (correlation and regression) (51.6%); NN (14.5%); DT (14.5%); Bayesian-based models (8%); SVM (3.2%);	Hybrid random Forest (99.25-99.98%); NN (98.81%); Random	Almost 86% of the synthesized models fall within the statistical modeling and supervised machine learning.

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

Author(s)	*Country*	*Review type*	*Time span*	*Number of articles reviewed*	*Subject areas/ Learning contexts*	*Total sample size*	*Purpose of review*	*Input (Predictor) variables*	*Output (Predicted) variables*	*EDM methods*	*EDM algorithms*	*EDM algorithms with highest prediction accuracy*	*Summary of the Results*[1]
					sciences/humanities (13%); & mix (8%)	kindergarten (1.61%). NB: 100 to >100,000 students	academic performance prediction by focusing on the attainment of learning outcomes.		perceived competence & achievement (n 5); self-reports about educational aspects (n = 3); failure, dropout or graduate rates (n = 3); other (e.g., college enrolment, career, etc.) (n = 6); NS (n = 1)		instance-based models (1.6%); & other (6.5%)	Forest (98%); NB (96.87%); & ANN (95.16-97.30/50	Regression, neural network, and tree-based models were the most used classification techniques for predicting the attainment of student learning outcomes.

Note. [1] = Summary of the results as they relate to the main focus of the current; N/M = Not mentioned, [2] = This excludes papers that did not provide specific number of students, [3] = total sample size of 15 articles only

Abbreviations: AI= artificial intelligence; ANN =; artificial neural networks; BC = Bayes classifiers; CGPA = cumulative grade point average; DM = data mining; EDM = educational data mining; DT = Decision tree; EM = Expectation maximisation; GUHA = general unary hypotheses automation; HE = higher education; KLSI = Klob Learning Style Inventory; K-NN = K-Nearest Neighbour; LA = learning analytics; LMS = learning management system; MLP = multi-layer perceptron; MOOCs = massive open online courses; NLP = natural language processing; NN = neural networks; RF = random forest; RMSE = Root mean Square error; SRMR = Standardized root mean square residual ; SAP = student academic performance; SAS EM = Statistical Analysis System Enterprise Miner; SNA = social network analysis; SPA; sequential pattern mining; SPSS = Statistical Package for the Social Sciences; SVM = support vector machines; VLE = virtual learning environment; WEKA = Waikato Environment for Knowledge Analysis