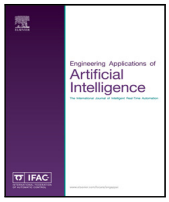
Engineering Applications of Artificial Intelligence 122 (2023) 106071

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Survey paper

Classification Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining — A survey

Sunita M. Dol [a,*], Pradip M. Jawandhiya [b]

[a] Department of Computer Science and Engineering, Walchand Institute of Technology, Solapur 413006, Maharashtra, India
[b] Department of Computer Science and Engineering, Pankaj Laddhad Institute of Technology and Management Studies, Yelgaon, Buldhana 443002, Maharashtra, India

ARTICLE INFO

Keywords:
Data mining
Educational data mining
Classification
Clustering
Association rule mining

ABSTRACT

Educational data mining (EDM) is the application of data mining in the educational field. EDM is used to classify, analyze, and predict the students' academic performance, and students' dropout rate, as well as instructors'performance in order to improve teaching–learning process. This review article discusses the detailed analysis of 142 research articles from publication year 2010-2020 downloaded from the research databases such as IEEE, Springer, ACM, and Elsevier. Also this review article contains the current happenings related to EDM in year 2021 and 2022. In this review article, the use of classification techniques and classification techniques along with other data mining techniques such as clustering algorithm, association rule algorithms, regression techniques and ensemble techniques in EDM are presented thoroughly. The comparative study is considered for Classification Techniques; Classification and Clustering Technique; Classification ans Association Rule Mining; Classification, Clustering and Association rule mining; Classification, Regression, and Clustering; and Classification, and Ensemble. Analysis in terms of Yearwise Number of Research Articles employing Classification Techniquein EDM; Classification with other Data Mining Technique used in EDM; classifier as per Weka Tool; Classification Techniques; Clustering Techniques; Association Rule Techniques; Selecting the best Classification Technique; Classification performance metric; software used in EDM; Sampling Period; size of dataset; and data mining tools are illustrated.

From review of 142 research articles, it is noted that classification techniques are mostly used technique for analyzing students' performance in EDM. Also classification technique along with clustering techniques are applied to predict the performance of students. It is found that Naïve Bays, Random Forest, Support vector machine and J48 are mostly considered classification techniques while in classification along with clustering techniques, K-means clustering algorithm is used with classification algorithms. The classification algorithms such as Naïve Bays, Random Forest and Support Vector Machine are noted to be the best classification algorithms after comparing various classification algorithms based on various performance parameters. Among various performance parameters, the parameters accuracy, precision, recall, f-measures and k-fold value found to be used by most of the research articles. Programming languages used to build the model in EDM for analyzing the students' dataset from educational setting, are Java, R and Python programming languages while data mining tools considered to evaluate the performance of classification or clustering or association rule algorithms are Weka, and RapidMiner. Classification algorithms under the classifiers as per Weka tool such as Tree, Bays, Function and PMML classifier are applied in most of the research articles.

In addition to comparative analysis and analysis based on various factors, research gaps are also identified and mentioned the same in this article. Future direction for researcher working in EDM related to building the model on the dataset obtained from educational setting to predict students' performance are discussed so that work in EDM can be carried out to improve the teaching–learning process.

Abbreviations: EDM, Educational Data Mining; LMS, Learning Management System; MOOC, Massive Open Online Course; LDA, Linear Discriminant Analysis; MLP, Multilayer Perceptron; ANN, Artificial Neural Network; SMO, Sequential minimal optimization; RBF, Radial Basis Function; KNN, K-Nearest Neighbour; SVM, Support Vector Machine; TP, True Positive; TN, True Negative; FP, False Positive; FN, False Negative; RMSE, Root Mean Square Error; CART, Classification And Regression Trees; ROC, Receiver Operating Characteristic; MAE, Mean Absolute Error; AUC, Area under the ROC Curve; CNN, Convolutional Neural Network; DT, Decision Tree; VQNN, Vaguely Quantified Nearest Neighbour; GPRT, Gaussian Processes Random Tree; SVD, Singular Value Decomposition

* Corresponding author.
E-mail addresses: sunita_aher@yahoo.com (S.M. Dol), pmjawandhiya@gmail.com (P.M. Jawandhiya).

https://doi.org/10.1016/j.engappai.2023.106071
Received 4 April 2022; Received in revised form 28 September 2022; Accepted 26 February 2023
Available online 11 March 2023
0952-1976/

1. Introduction

Educational data mining is nothing but collecting the data from educational setting, classifying this data as per requirement, and analyzing data for predicting academic performance of students, identifying the factors in early stages that leads to the students' dropout, predicting students' performance in various programming courses, etc. An educational system has huge amount of data which includes students' academic record, students' placement data, students' performance in programming languages, log data, instructors' data, etc. This data is from various resources such as online courses, Learning Management System (LMS)/ Course Management System (CMS) such as MOODLE (Modular Object-Oriented Dynamic Learning Environment) data, educational software, etc. Data Mining techniques are applied on this educational data to extract meaningful and useful pattern which can further be considered to improve the teaching–learning process. Various data mining techniques such as classification, clustering, regression, neural network, genetic algorithms, association rule algorithms, etc. are used to discover the knowledge from this educational dataset. In this direction, various models have been proposed to analyze and predict students' performance in academic as well as in onlines learning. Such six models are identified such as predicting the dropout rate, predicting the students' performance in online education/ courses, analyzing students' performance in programming courses, analyzing the Learning Management System (LMS) data, predicting the students' academic performance and predicting the students' placement

First model is to predict the dropout rate. School/college failure/dropout is the major problem that institute faces in education sector. There are many research articles Márquez-Vera et al. (2013b), Palazuelos et al. (2013), Manhães et al. (2014), Barbosa Manhães et al. (2015), Meedech et al. (2016), Niu et al. (2018), Pérez et al. (2018), Burgos et al. (2018), Tasnim et al. (2019), Chango et al. (2019), Tarmizi et al. (2019), Vila et al. (2018), Lottering et al. (2020), M.Á. et al. (2020), Santoso (2019), and (Ribeiro and Canedo, 2020) that deals with analyzing the students' behavior and reason for dropout. In research articles Márquez-Vera et al. (2013b), Palazuelos et al. (2013), Manhães et al. (2014), Meedech et al. (2016), Niu et al. (2018), Pérez et al. (2018), Burgos et al. (2018), Tasnim et al. (2019), Chango et al. (2019), Tarmizi et al. (2019), Vila et al. (2018) , Lottering et al. (2020), Mkwazu and Yan (2020), and Ribeiro and Canedo (2020), author/s compared various classification techniques and suggested the best classification technique based on various performance metrics while the research articles M.Á. et al. (2020), Santoso (2019) considered the classification as well as clustering algorithm to predict the dropout rate.

Second model is to predict the students' performance in online education/ courses. The research article Ayub et al. (2017), Sukhija et al. (2018) demonstrated the Classification and Association rules techniques to analyze the educational dataset while the research articles Cocea and Weibelzahl (2010), Wang and Liao (2011), Nasiri et al. (2012), Shukor et al. (2015), Buenaño-Fernández et al. (2017), Figueira (2017), Costa et al. (2017), Abyaa et al. (2018), Al Fanah and Ansari (2019), Ma et al. (2021), Teoh et al. (2022), and Vinker and Rubinstein (2022) suggested the classification techniques to predict the performance of students in online mode. Authors Angra and Ahuja (2017) focused on three data mining techniques — Classification, Linear Regression, and Clustering to understand the behavior and performance of students using data mining tool RapidMiner while the authors (Bodea et al., 2010) presented the classification and clustering technique to analyze students' performance in Project Management course with the help of data mining tool WEKA.

Third model is to do the analysis of students' performance in programming courses. The research articles Márquez-Vera et al. (2013a), Pathan et al. (2014), Ahadi et al. (2015), Sisovic et al. (2015), Amornsinlaphachai (2016), Badr et al. (2016), Leppänen et al. (2017), Costa et al. (2017), and Lagus et al. (2018) deals with studying the performance of students in programming course. The research article Badr et al. (2016) analyzed the programming course related data using Classification and Association Rule Mining to predict the performance in course based on Mathematics and English course. The research articles Márquez-Vera et al. (2013a), Pathan et al. (2014), Ahadi et al. (2015), Sisovic et al. (2015), Amornsinlaphachai (2016), Leppänen et al. (2017), Costa et al. (2017), and Lagus et al. (2018) used the classification technique for programming course result analysis based on various performance metrics.

Fourth model is to analyze the Learning Management System (LMS) data. The authors (Kan et al., 2010) considered the data mining techniques Association Rule, Classification and Clustering to predict students' performance in the course Decision Analysis of Information Management department conducted on LMS.

Fifth model is to predict the students' academic performance. The research articles Sakurai et al. (2012), Bresfelean et al. (2012), Hoe et al. (2013), Chau and Phung (2013), Göker et al. (2013), Mashiloane and Mchunu (2013), Anh et al. (2014), Auddy and Mukhopadhyay (2015), Devasia et al. (2016), Hamsa et al. (2016), Malini and Kalpana (2021), Dabhade et al. (2021), Xu et al. (2021), Veluri et al. (2022), Hussain et al. (2022) used data mining techniques such as classification, clustering and association rule algorithm to analyze the academic record and predict the students' academic performance in education sector based various performance metrics of various data mining algorithms.

Sixth model to predict the students' placement. In the research articles Şen et al. (2012), Pratiwi (2013), Gera and Goel (2015), Pruthi and Bhatia (2015), and Ramanathan et al. (2016), data mining technique — such as classification technique is used and evaluated based on the performance metric to predict the students' eligibility and performance for placement in higher education.

In last two years 2021 and 2022, research related to EDM used the dataset from various repositories and educational institutes to analyze the students' performance using various visualization tools, data mining tools, and data mining techniques. Following are some of the findings from these years-

- In research Hernández-Leal et al. (2021), authors performed the descriptive and exploratory analysis of the data collected from educational institutes using the Tableau tool to obtain information about students' behavior while decision tree classification algorithms using Orange & Weka and Clustering algorithms using the RapidMiner are used to find the educational pattern.
- Student performance dataset from UCI is analyzed using various classifiers to predict the various factors that affects the student performance (Malini and Kalpana, 2021; Veluri et al., 2022; Hussain et al., 2022).
- The classification algorithms such as multiple linear regression and support vector regression are considered to analyze the questionnaire data based on personal, educational, behavioral and extra-curricular using Python3 tool (Dabhade et al., 2021).
- Use of classification algorithms such as JRip, PART, ZeroR, Naïve Bayes, and J48 and K-means clustering algorithm to address the academic procrastination is mentioned in the research article Xu et al. (2021).
- Attention predicting model (APM) algorithm is proposed and compared with various classification algorithms such as Logistic regression, Naive Bayes, DecisionStump, JRip, J48 and K-nearest neighbour are compared on various performance parameter Accuracy, Precision, Recall, F-Measure, Kappa statistic, MSE, and RMSE to find the students' performance in e-learning (Ma et al., 2021).
- Classification algorithms are also used to predict student performance in MOOC (Teoh et al., 2022; Vinker and Rubinstein, 2022).

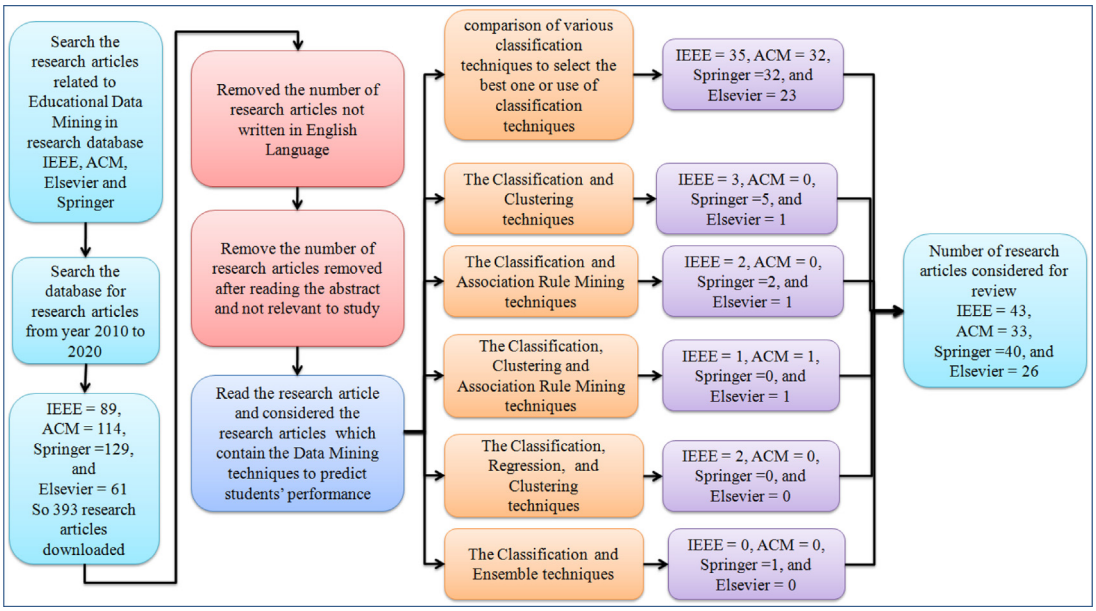


Fig. 1. Review process.

In this review article, 142 research articles from IEEE, ACM, Elsevier and Springer research databases for the publication year 2010–2020 are used. The analysis is done based on number of articles related to various data mining techniques in various research database, Publication Year of research articles, research database, Number of Citations, and Number Pages in Research Article. The use of classification techniques such as Naïve Bays, Random Forests, Support vector machine, etc. and classification techniques along with other data mining techniques such as clustering algorithm, association rule algorithms, regression techniques and ensemble techniques in EDM are elaborated to analyze students' performance. The comparative study is considered for Classification Techniques along with its' combination with other data mining techniques. Analysis in terms of Yearwise Number of Research Articles employing Classification Techniquein EDM; Classification with other Data Mining Technique used in EDM; classifier as per Weka Tool; Classification Techniques; Clustering Techniques; Association Rule Techniques; Selecting the best Classification Technique; Classification performance metric; software used in EDM; Sampling Period; dataset size; and data mining tools are illustrated. Analysis of research articles based on performance parameters such as Analysis on basis of Accuracy, TP rate, FP rate, Precision, Recall, F-measures, Cross Validation, Correctly Classified Instance, Incorrectly Classified Instance, RMSE (Root mean square error), MAE (Mean Absolute Error), AUC (Area Under the Curve), and ROC (Receiver operating characteristic curve) are also mentioned in this review article.

The article is arranged in the following manner Section 2 describe systematic review method for selecting the research articles and Section 3 elucidates EDM and comparative study. Section 4 describe analysis and discussion, while Section 5 provides the research gap and future direction followed by conclusion in Section 6.

2. Systematic review method and analysis

This section discusses about the systematic review method and analysis of research articles based on publication year, number of citation, research database, etc. Following steps are considered to search the research articles for this review paper -

- Since this review article is related to use of classification techniques and its Combination with Clustering and Association Rule Mining in Education Data Mining, so all research articles related to the keyword "Educational Data Mining" were searched from various research databases such as IEEE, ACM, Elsevier and Springer. So all journals as well as proceeding research articles were downloaded from these research databases.
- To download the research articles from research database, the first filter applied was publication year that was from 2010 to 2020.
- So such 393 research articles were downloaded. Out of 393 research articles, 89 from IEEE, 114 from ACM, 129 from Springer and 61 from Elsevier research database were considered.
- After going through these research articles, some research articles were not in English language. So such articles were removed from the list.
- Those articles which are not relevant to the study in EDM after going through the abstract, were removed.
- From remaining research articles, only those research articles which contained the analysis of students' performance using Data Mining techniques such as classification, clustering and association rule mining algorithms were considered for review purpose. The research articles were classified into following categories
  - Comparison of various Classification algorithms to select best one or use of Classification algorithm;
  - Classification and Clustering algorithms;
  - Classification and Association Rule Mining algorithms;
  - Classification, Clustering and Association Rule Mining algorithms;
  - Classification, Regression, and Clustering algorithms; and
  - Classification and Ensemble techniques algorithms.
- Finally 142 research articles were considered out of which 43 research articles were from IEEE, 33 from ACM, 40 from Springer and 26 from Elsevier as shown in Fig. 1.

2.1. Analysis based on number of articles related to various data mining techniques in various research database

This subsection deals with the analysis on basis of the number of articles related to various data mining techniques downloaded from various research database. All selected research articles for review purpose are categorized into six groups as shown in Table 1:

- Classification algorithms — 35 research articles from IEEE, 32 from ACM, 32 from Springer and 23 from Elsevier so total 142

Table 1
Analysis based on number of articles related to various data mining techniques in various research database.

Technique	IEEE	ACM	Springer	Elsevier	Total
Classification	35	32	32	23	122
Classification and Clustering	3	0	5	1	9
Classification and Association Rule Mining	2	0	2	1	5
Classification, Clustering and Association Rule Mining	1	1	0	1	3
Classification, Regression, and Clustering	2	0	0	0	2
Classification and Ensemble techniques	0	0	1	0	1

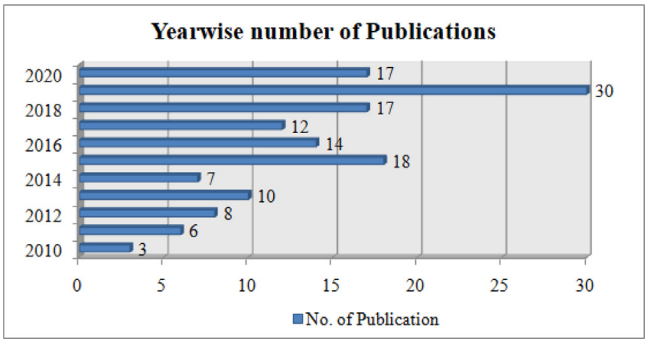


Fig. 2. Yearwise number of research articles considered for review process.

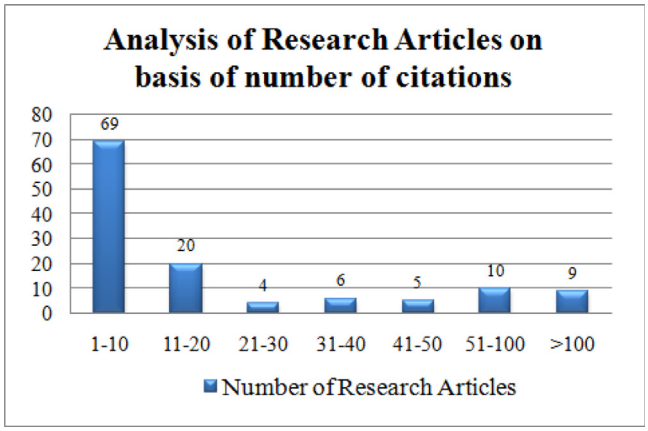


Fig. 3. Analysis based on number of citations.

research articles are considered which used the classification algorithms for analyzing students' performance.

- Classification and Clustering algorithms — 3 research articles from IEEE, 5 from Springer and 1 from Elsevier so total 9 research articles made the use of both classification and clustering algorithms.
- Classification and Association Rule Mining algorithms — 2 research articles from IEEE, 2 from Springer and 1 from Elsevier so total 5 research articles applied both classification and association rule algorithms.
- Classification, Clustering and Association Rule Mining algorithms — 1 research article from IEEE, 1 from Springer and 1 from Elsevier so total 3 research articles considered three Data Mining algorithms that is classification, Clustering, and association rule algorithms for determining students' performance.
- Classification, Regression, and Clustering algorithms — 2 research articles from IEEE used three algorithms Classification, Regression, and Clustering algorithms to predict the students' performance in EDM.
- Classification and Ensemble techniques algorithms — Two algorithms such as Classification and Ensemble techniques are applied by only one research article from Springer.

2.2. Analysis based on publication year

This section describes the analysis of research articles based on the publication year. Seventeen research articles are considered from publication year 2020 and 2018 while 30 research articles are from year 2018. Number of research articles downloaded for publication year 2015, 2016, and 2017 are 18, 14, and 12 respectively. Very few research articles are from year 2010 that is three research articles (see Fig. 2).

2.3. Analysis on basis of publication year and research database

This subsection discusses the analysis of research articles based on publication year as well as research database from which it is accessed. Table 2 presents the analysis of research articles based on publication year from 2010 to 2020 and research database IEEE, ACM, Springer and Elsevier. It is noted from Table 2 that out of 142 research articles, 43 from IEEE, 33 from ACM, 40 from Springer and 26 from Elsevier are considered for this review article. Maximum research articles considered are from publication year 2019 while very few research articles are from 2010.

2.4. Analysis on basis of number of citations

Number of Citation plays an important role in deciding the number of researcher referred that particular research article. Fig. 3 shows the analysis of research articles based on the number of citation to that particular research article in the range 1–10, 11–20, 21–30, 31–40, 41–50, 51–100, and >100. From Fig. 3, it is noted that 20 research articles have the citation in the range 11–20 while 69 research articles have the citation in the range 1–10. There are 10 research articles having number of citations in the range 51–100 while the number of citations for 9 research articles are more than 100.

2.5. Analysis on basis of number pages in research article

Detailed analysis of problem statement related to the EDM in research articles depends on number of pages in it. Fig. 4 presents the analysis based on number of pages in research articles with eight ranges 2–4, 5–7, 8–10, 11–13, 14–16, 17–19, 20–22 and 23–25. Fifty seven research articles are in the page range 5–7 while twenty nine research articles contains the number of pages in the range 8–10. The research articles Dejaeger et al. (2012), Márquez-Vera et al. (2013a), Chen et al. (2014), Mayilvaganan and Kalpanadevi (2015), Stahovich and Lin (2016), Kassak et al. (2016), Pérez et al. (2018), Burgos et al. (2018), Miguéis et al. (2018), Francis and Babu (2019), Adekitan and Salau (2020), Crivei et al. (2019), Dimić et al. (2019), Almutairi et al. (2019), Santoso (2019), El Aissaoui et al. (2020), Agrawal et al. (2020), and Injadat et al. (2020b) consists of pages in the range 14–16 while the research articles Cocea and Weibelzahl (2010), Mashiloane and Mchunu (2013), Jishan et al. (2015) , Salinas and Stephens (2015), Ramanathan et al. (2016), Meedech et al. (2016),

Table 2
Analysis based on publication year and research database.

Year	IEEE		ACM		Springer			Elsevier		Total
	Number of Journal articles	Number of proceeding articles	Number of Journal articles	Number of proceeding articles	Number of Journal articles	Number of Book chapters	Number of proceeding articles	Number of Journal articles	Number of proceeding articles	
2010	0	1	0	0	0	1	0	1	0	**3**
2011	1	1	0	0	0	0	2	1	1	**6**
2012	0	3	0	1	0	1	0	2	1	**8**
2013	1	5	0	0	1	1	1	1	0	**10**
2014	1	3	0	2	0	0	0	1	0	**7**
2015	1	5	0	5	0	1	2	0	4	**18**
2016	1	3	0	2	0	2	0	2	4	**14**
2017	0	3	0	6	1	0	1	1	0	**12**
2018	0	4	1	6	0	1	2	2	1	**17**
2019	1	5	0	7	2	3	10	2	0	**30**
2020	2	2	0	3	3	2	3	1	1	**17**
Total	**8**	**35**	**1**	**32**	**7**	**12**	**21**	**14**	**12**	**142**

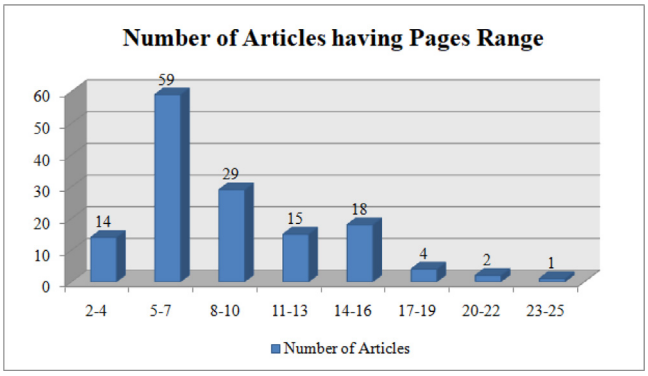


Fig. 4. Analysis based on number of pages in research articles.

Maitra et al. (2018), Akram et al. (2019), Tarmizi et al. (2019), Martins et al. (2019), Kasthuriarachchi and Liyanage (2018), Ibrahim et al. (2018), Karthikeyan et al. (2020), Ajibade et al. (2018), and Ashraf et al. (2020) have the page range 11–13. Fourteen research articles Kan et al. (2010), Sakurai et al. (2012), Hoe et al. (2013), Pratiwi (2013), Dangi and Srivastava (2014), Manhães et al. (2014), Gera and Goel (2015), Guo et al. (2015), Lehr et al. (2016), Sanchez-Santillan et al. (2016), Kitanaka et al. (2017), Patil et al. (2018), Rojanavasu (2019), Islam and Mahmud (2020) have the page range in 2–4 pages. The research articles Pise and Kulkarni (2017), Lagus et al. (2018), Yahya (2019), M.Á. et al. (2020) - number of pages in the range 17–19 Yousafzai et al. (2020), Adekitan and Salau (2019) -the number of pages in the range 20–22 and Injadat et al. (2020a) - the number of pages in the range 23–35.

3. Educational data mining

Educational Data Mining is the application of Data Mining (DM) in which DM techniques are applied on the dataset obtained from educational setting to analyze the students' performance for helping the institutes to improve the teaching–learning process. Generally used data mining techniques in EDM are classification, clustering and association rule algorithm.

Classification is a supervised technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category/class to which a new data will fall under. It is also called supervised learning. Example: Classification of emails as "spam" or "not spam" depending on header, sender, and content etc.

Classification is a data mining task that maps the data into predefined groups & classes. It is also called as supervised learning Two data sets are required for classification technique-

- Training Set — Given a collection of records in which each record contains a set of attributes, one of the attributes is the class.
- Test Set — is used to determine the accuracy of the model.

Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it. It consists of two steps: Model construction and Model usage.

Model Construction — It consists of set of predetermined classes. Each tuple/sample is assumed to belong to a predefined class. The set of tuple used for model construction is training set. The model is represented as classification rules, decision trees, or mathematical formulae.

Model usage — This model is used for classifying future or unknown objects. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set, otherwise over-fitting will occur.

Clustering is an unsupervised learning and divides the dataset into number of clusters such that data belonging to a cluster have same characteristic. Types of clustering are Hierarchical clustering, Partitioning methods, Density based clustering, constraint based clustering, Fuzzy clustering, and Distribution based clustering.

Association rule mining is the unsupervised learning which has two parts — antecedent and consequent. It is rule-based algorithm used to find the relationship between the variables in the database, e.g. if a person buy the mobile then he is likely to buy the mobile cover and screenguard for the same. Different association rule algorithms are Apriori association rule, Filtered growth, relational association rule mining, etc.

Fig. 5 shows the process for determining students'performance. First step is to collect the data from educational institutes such students' academic record, background data, data from Learning Management System, etc. Educational data is collected for particular period that is number of years, semester data, etc. Collected data may contain the outliers, missing values, etc. So the data needs to be preprocessed. The preprocessing involves missing value treatment, Outlier, finding out the number of features and their relationship with each other so that during the training of the model the accuracy of result should be improved. The process of extracting the most continuous, non-redundant and relevant characteristics to employ in model creation is known as feature selection. As the number and diversity of datasets grow, it is more critical than ever to reduce them methodically. After feature extraction step, data is prepared in the format required for building the model. Building the model for analysis of data is the next step.

So here for building the model, various possibilities are

1. **Comparison of classification techniques** – in this case various classification algorithms such as Naïve Bays, Random Forests, Support Vector Machine, etc. are compared based on various

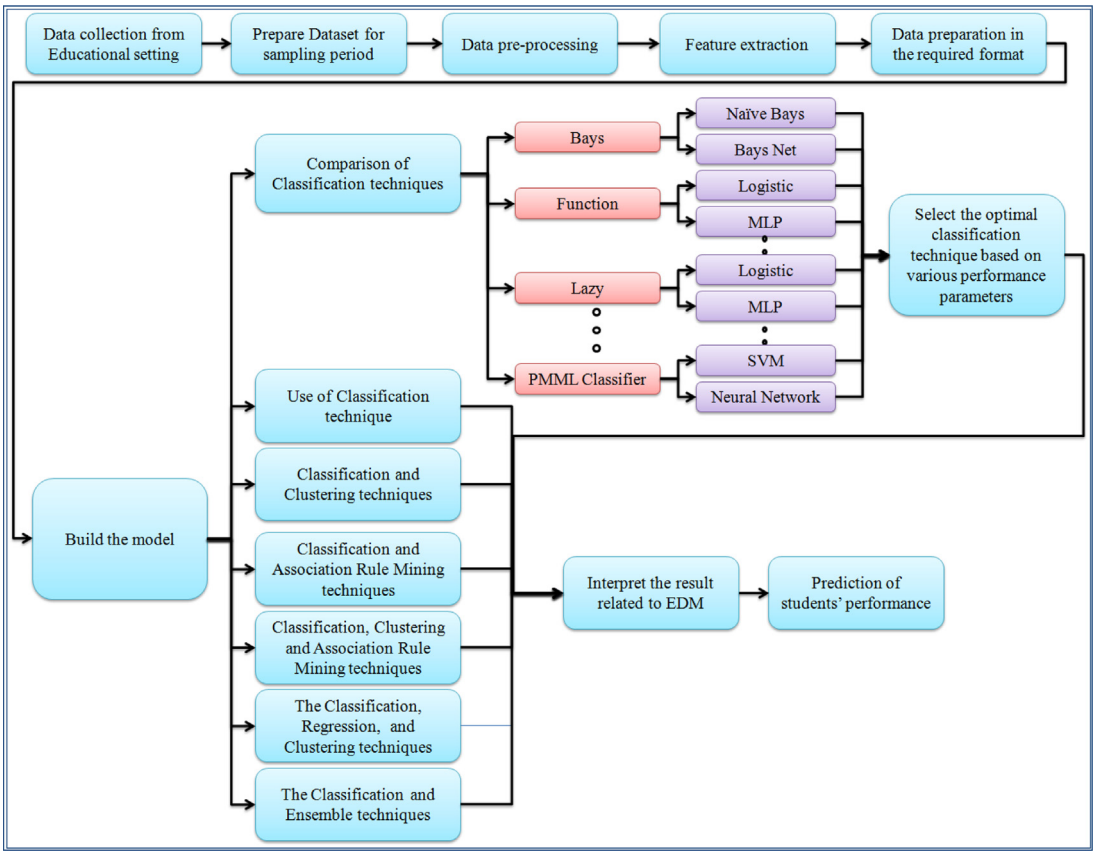


Fig. 5. Data Mining process to predict students' performance.

performance parameters such as Accuracy, Precision, Recall, F-measures, etc. to find the best classification algorithm that work for given dataset.

2. **Use of Classification techniques** — here classification algorithm is applied to the dataset and performance of the algorithm is measured based various performance metrics.
3. **Use of Classification and Clustering techniques** – in this case, both – classification algorithms considered steps in 1 or 2 as well as clustering algorithms such as K-means, Expectation Maximization, etc. are applied on the dataset.
4. **Use of Classification and Association Rule Mining techniques** — here, both classification algorithms considered steps in 1 or 2 as well as association rule algorithms such as Apriori association rules, FPGrowth, etc. are considered on the dataset.
5. **Use of Classification, Clustering and Association Rule Mining techniques** — here, three Data Mining techniques such as classification algorithms considered steps in 1 or 2, clustering algorithms and association rule algorithms are used on the dataset.
6. **Use of Classification, Regression, and Clustering techniques** — in this case, three techniques such as classification algorithms considered steps in 1 or 2, regression algorithms such as Linear Regression, Multiple Regression, etc. and clustering algorithms are applied on the dataset.
7. **Use of Classification and Ensemble technique techniques** — here, classification algorithms considered steps in 1 or 2, as well as ensemble techniques such as Adaboost, Bagging, etc. are used on the dataset.

After building the model using above mentioned possibilities, results are obtained and interpreted to predict the performance of students.

Fig. 6 shows the classification of Data Mining techniques that can be used in EDM. There are basically three data mining techniques that is classification, clustering and association rule. Classification algorithms are further classified as Bays, Function, Lazy, Meta, Rules, Tree, and PMML (Predictive Model Markup Language) classifier based on the DM tool Weka. Clustering techniques are classified as non-hierarchical and hierarchical techniques.

Table 3 shows the analysis of research articles on basis of Data Mining techniques such as classification, clustering and association rule. Table 3 contains the following techniques along with the reference number of research articles referring those techniques-

- Classification
- Classification and Clustering
- Classification and Association Rule Mining
- Classification, Clustering and Association Rule Mining
- Classification, Regression, and Clustering
- Classification and Ensemble technique

It is noted from Table 3 that 122 research articles referred various classification algorithms for analyzing students' performance. The research articles Bodea et al. (2010), Chellatamilan et al. (2011), Zengin et al. (2011), Bresfelean et al. (2012), Akram et al. (2019), Francis and Babu (2019), M.Á. et al. (2020), Santoso (2019), El Aissaoui et al. (2020) used the data mining techniques Classification and Clustering while the research articles Badr et al. (2016), Ayub et al. (2017), Sukhija et al. (2018), Rojanavasu (2019), Crivei et al. (2019) considered Classification and Association Rule Mining. The research articles Kan et al. (2010), Trandafili et al. (2012), Dejaeger et al. (2012) applied three techniques Classification, Clustering and Association Rule Mining on students' dataset whereas the papers Jacob et al. (2015), Angra and Ahuja (2017) made use of Classification, Regression, and Clustering. The research article Ajibade et al. (2018) used two techniques that is the Classification and Ensemble techniques.

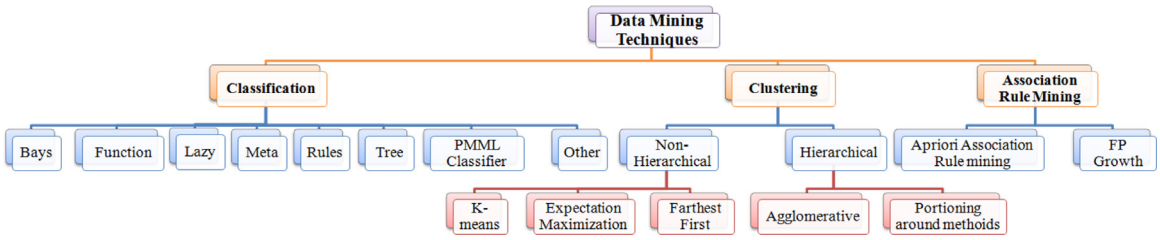


Fig. 6. Data Mining techniques.

Table 3
Analysis of research articles on basis of DM techniques.

Techniques	Number of research articles	Reference number
Classification	122	Guruler et al. (2010), Cocea and Weibelzahl (2010), Chuan et al. (2011), El-Halees (2011), Wang and Liao (2011), Sakurai et al. (2012), Bunkar et al. (2012), Nasiri et al. (2012), Dejaeger et al. (2012), Şen et al. (2012), Sen and Ucar (2012), Abaya and Gerardo (2013), Hoe et al. (2013), Chau and Phung (2013), Pratiwi (2013), Göker et al. (2013), Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Mashiloane and Mchunu (2013), Palazuelos et al. (2013), Blagojević and Micić (2013), Dangi and Srivastava (2014), Pathan et al. (2014), Anh et al. (2014), Chen et al. (2014), Ragab et al. (2014), Manhães et al. (2014), Natek and Zwilling (2014), Gera and Goel (2015), Pruthi and Bhatia (2015), Parmar et al. (2015), Guo et al. (2015), Guarín et al. (2015), Sorour et al. (2015), Ahadi et al. (2015), Sisovic et al. (2015), Bakaric et al. (2015), Barbosa Manhães et al. (2015), Jishan et al. (2015), Salinas and Stephens (2015), Auddy and Mukhopadhyay (2015), Shukor et al. (2015), Kaur et al. (2015), Amornsinlaphachai (2015), Amornsinlaphachai (2016), Devasia et al. (2016), Lehr et al. (2016), Agaoglu (2016), Chaudhury et al. (2016), Sanchez-Santillan et al. (2016), Ramanathan et al. (2016), Meedech et al. (2016), Stahovich and Lin (2016), Hassan and Al-Razgan (2016), Hamsa et al. (2016), Ahmed et al. (2016), Kassak et al. (2016), Athani et al. (2017), Castro-Wunsch et al. (2017), Buenaño-Fernández et al. (2017), Figueira (2017), Kitanaka et al. (2017), Daud et al. (2017), Leppänen et al. (2017), Pise and Kulkarni (2017), Jung (2016), Costa et al. (2017), Kiu (2018), Srivastava et al. (2018), Patil et al. (2018), Kaunang and Rotikan (2018), Chitra and Agrawal (2019), Lagus et al. (2018), Spatiotis et al. (2018), Niu et al. (2018), Rustia et al. (2018), Abyaa et al. (2018), Chanlekha and Niramitranon (2018), Martínez-Abad et al. (2018), Kularbphettong (2017), Pérez et al. (2018), Burgos et al. (2018), Miguéis et al. (2018), Maitra et al. (2018), Tasnim et al. (2019), Ketui et al. (2019), Al Breiki et al. (2019), Alshehri (2019), Lagman et al. (2019), Umer et al. (2019), Amazona and Hernandez (2019), Chango et al. (2019), Altaf et al. (2019), Al Fanah and Ansari (2019), Adekitan and Salau (2020), Tarmizi et al. (2019), Martins et al. (2019), Rawat and Malhan (2019), Kamal and Ahuja (2019), Dimić et al. (2019), Zaffar et al. (2018), Vila et al. (2018), Almutairi et al. (2019), Kasthuriarachchi and Liyanage (2018), Ab Ghani et al. (2018), Ibrahim et al. (2018), Yousafzai et al. (2020), Adekitan and Salau (2019), Yahya (2019), Lottering et al. (2020), Utari et al. (2020), Mengash (2020), Rahman and Mahmud (2020), Mkwazu and Yan (2020), Islam and Mahmud (2020), Injadat et al. (2020a), Karthikeyan et al. (2020), Jung (2018), Agrawal et al. (2020), Ribeiro and Canedo (2020), Ashraf et al. (2020), Injadat et al. (2020b)
Classification and Clustering	9	Bodea et al. (2010), Chellatamilan et al. (2011), Zengin et al. (2011), Bresfelean et al. (2012), Akram et al. (2019), Francis and Babu (2019), M.Á. et al. (2020), Santoso (2019), El Aissaoui et al. (2020)
Classification and Association Rule	5	Badr et al. (2016), Ayub et al. (2017), Sukhija et al. (2018), Rojanavasu (2019), Crivei et al. (2019)
Classification, Clustering and Association Rule Mining	3	Kan et al. (2010), Trandafili et al. (2012), Dejaeger et al. (2012)
Classification, Regression, and Clustering	2	Jacob et al. (2015), Angra and Ahuja (2017)
Classification and Ensemble techniques	1	Ajibade et al. (2018)

3.1. Feature selection and extraction

In Data Mining, Feature Selection and Extraction are two important aspects to be considered. So first step is the feature extraction and then on obtained data, feature selection methods are applied. In Feature extraction, the raw data is transformed into the numerical features and these numerical features are processed without changing the information in original data. Feature extraction can be unsupervised such as Principal Component Analysis (PCA) or supervised such as Linear Discriminant Analysis (LDA).

Feature selection is use to filter out the irrelevant data from dataset Feature selection can also be unsupervised such as Variance Thresholds or supervised such as Genetic Algorithms. Multiple methods can be combined if required. The difference between Feature Selection and Extraction is that feature extraction creates new dataset while feature selection keeps the subset of original features. Unsupervised feature selection is used to reduce features in text clustering process which are not useful (Abualigah et al., 2016; Abualigah and Khader, 2017). To solve the problem of feature selection in text document clustering, the technique Feature selection based on harmony search algorithm is suggested by authors (Abualigah et al., 2018).

Table 4 gives the analysis of research articles based on feature selection/ extraction methods. Feature extraction methods considered are Principal Component Analysis (PCA), Sixth order Butterworth band pass filter, and Common spatial patterns (CSP) spatial filter.

Feature selection methods used for analysis are Chi-square, Information gain, OneR, Fuzzy rough feature selection, Correlation coefficient, Feature selection using Support Vector Machine, AdaBoost classifier, Tree Ensemble, Histogram Discretization, Fast Correlation Based Feature Selection, K-best feature selection, Krippendorf's Alpha measure, Feature selection using Genetic Algorithm, Forward feature selection, Filter approach based on Term Frequency, Random feature selection, Linear Regression technique, Exploratory data analysis, Discretization and Resampling, Markov Blanket, Wrapper feature selection, Feature selection method using the artificial neural network, Modified Chameleon swarm algorithm, Minimum redundancy maximum relevancy and Feature importance ranking method (FIRM). Feature selection methods such as SymmetricalUncertAttributeEval,

Table 4
Analysis of research articles on basis of feature selection/ extraction methods.

Feature Selection/ Extraction Method	Ref.
Feature Extraction	
Principal Component Analysis (PCA)	M.Á. et al. (2020), Injadat et al. (2020b)
Sixth order Butterworth band pass filter	Kapgate (2022)
Common spatial patterns (CSP) spatial filter	Kapgate (2022)
Feature Selection Method	
Chi-square	Cocea and Weibelzahl (2010), Zaffar et al. (2018)
Information gain	Cocea and Weibelzahl (2010), Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Ahadi et al. (2015), Bakaric et al. (2015), Stahovich and Lin (2016), Daud et al. (2017), Costa et al. (2017), Tasnim et al. (2019), Al Breiki et al. (2019), Akram et al. (2019), Tarmizi et al. (2019), Almutairi et al. (2019), Ajibade et al. (2018)
OneR	Cocea and Weibelzahl (2010)
SymmetricalUncertAttributeEval	Göker et al. (2013), Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a)
Correlation-based feature subset selection	Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Ahadi et al. (2015)
Consistency-SubsetEval	Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a)
FilteredAttributeEval	Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a)
FilteredSubsetEval	Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a)
GainRatioAttributeEval	Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Daud et al. (2017), Mkwazu and Yan (2020)
ReliefFAttributeEval	Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Zaffar et al. (2018)
SymmetricalUncert-AttributeEval	Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a)
Fuzzy rough feature selection	Auddy and Mukhopadhyay (2015)
Correlation coefficient	Badr et al. (2016)
Feature selection using Support Vector Machine	Kitanaka et al. (2017)
AdaBoost classifier	Lagus et al. (2018)
Tree Ensemble	Niu et al. (2018)
Histogram Discretization	Dimić et al. (2019)
Fast Correlation Based Feature Selection	Zaffar et al. (2018)
K-best feature selection	Almutairi et al. (2019)
Krippendorf's Alpha measure	Ibrahim et al. (2018)
Feature selection using Genetic Algorithm	Yousafzai et al. (2020)
Forward feature selection	Adekitan and Salau (2019)
Filter approach based on Term Frequency	Yahya (2019)
Random feature selection	Utari et al. (2020)
Linear Regression technique	Mengash (2020)
Exploratory data analysis	Islam and Mahmud (2020), Santoso (2019)
Discretization and Resampling	Ajibade et al. (2018)
Markov Blanket	Jung (2018)
Wrapper feature selection	Abualigah and Diabat (2022)
Feature selection method using the artificial neural network	AlShourbaji et al. (2022)
Modified Chameleon swarm algorithm	Mostafa et al. (2022)
Minimum redundancy maximum relevancy	Alomari et al. (2017)
Feature importance ranking method (FIRM)	Rahman and Mahmud (2020)

Correlation-based feature subset selection, Consistency-SubsetEval, FilteredAttributeEval, FilteredSubsetEval, GainRatioAttributeEval, ReliefFAttributeEval, and SymmetricalUncertAttributeEval are considered from Weka tool.

From Table 4, it is found that feature extraction method Principal Component Analysis (PCA) is applied in research articles M.Á. et al. (2020), Injadat et al. (2020b) for extracting the features from dataset. In research article Kapgate (2022), Sixth order Butterworth band pass filter is used to filter the Event Related Potentials (P300) – acquired EEG signal while common spatial patterns (CSP) spatial filter is considered to minimize the discrimination among different conditions related to the Steady State Visual Evoked Potentials (SSVEP) - EEG signals. The author (Aldosari et al., 2022) suggested the use of the technique Generalized Normal Distribution Optimization (GND) and Dwarf Mongoose Optimization Algorithm (DMOA), followed by the Opposition-based Learning Mechanism (OBL) for feature extraction. It is also noted from Table 4 that Feature selection method - Information gain is used by research articles Cocea and Weibelzahl (2010), Márquez-Vera et al. (2013b,a), Ahadi et al. (2015), Bakaric et al. (2015), Stahovich and Lin (2016), Daud et al. (2017), Costa et al. (2017), Tasnim et al. (2019), Al Breiki et al. (2019), Akram et al. (2019), Tarmizi et al. (2019), Almutairi et al. (2019), and Ajibade et al. (2018) while GainRatioAttributeEval is considered in Márquez-Vera et al. (2013b,a), Daud et al. (2017), and Mkwazu and Yan (2020). The author (Wang et al., 2022) proposed the use of the technique Enhanced Remora Optimization Algorithm (ROA) for feature selection.

3.2. Classification techniques

Classification algorithm is used to classify the new data into the category based on certain characteristics. Example of classification algorithms are document classification, speech recognition, handwriting recognition, etc. Table 5 shows the classification techniques based on classifier such as Bays, Function, Lazy, Meta, Rules, Tree, and PMML Classifier considered in WEKA tool e.g. for Bays classifier, algorithms are - 1.Naïve Bayes and 2.Bays Net; Function classifier algorithms - 1. Logistic, 2. Simple Logistic, 3. Linear Discriminating analysis and its variations such as Probabilistic Multi-class Linear Discriminant Analysis classifier (Kapgate, 2022), 4. Multilayer Perceptron (MLP), 5. Artificial Neural Network, 6. Sequential minimal optimization (SMO), 7. Linear Regression, 8. Multiple Regression, 9.RBF Network; Lazy classifier algorithms - 1.IBk (Instance based Learner), 2.K-Nearest Neighbour (KNN), 3. Kstar, 4. QuickRules Fuzzy-Rough rule induction, 6. Fuzzy Nearest Neighbour, 7. Fuzzy Rough Nearest Neighbour and 8. Vaguely Quantified Nearest Neighbour (VQNN); etc. Table 4 also contains the other classification algorithms such as Decision Tree (not mentioned Decision Tree algorithm), Gradient boosted tree, Gaussian Processes, Random Tree (GPRT), Classification using CHAID (Chi-Squared Automatic Interaction Detection), Discriminant analysis, GLM: Generalized Linear Model, Singular Value Decomposition (SVD), Interpretable Classification Rule Mining, MaxMargin Multi-Label (M3L) classifier, LIBSVM, Cross-out classifier, Classifier based on polynomial regression and stochastic gradient descent, and Improved ID3.

Table 5
Classification techniques based on classifier considered in WEKA tool.

Bays	Function	Lazy	Meta
1. Naïve Bayes 2. Bays Net	1. Logistic 2. Simple Logistic 3. Linear Discriminating analysis 4. Multilayer Perceptron (MLP) 5. Artificial Neural Network 6. Sequential minimal optimization (SMO) 7. Linear Regression 8. Multiple Regression 9. RBF Network	1. IBk (Instance based Learner) 2. K-Nearest Neighbour (KNN) 3. Kstar 4. QuickRules Fuzzy-Rough rule induction 6. Fuzzy Nearest Neighbour 7. Fuzzy Rough Nearest Neighbour 8. Vaguely Quantified Nearest Neighbour (VQNN)	1. AdaBoost 2. Attribute Selected Classifier 3. Bagging 4. Classification Vis Regression 5. Logit Boot
Rules	**Tree**	**PMML Classifier**	**Other**
1. Decision Table 2. Conjuctive 3. Rules 4.Jrip 5. OneR 6. PART 7. Prism 8. RIDOR 9. Nnge	1. Decision Stump 2. J48 3. M5P 4. ADTree 5. Random Forest 6. Random Tree 7. REPTree 8.CART (Classification and Regression Trees) 9. SimpleCart 10.ID3 11.C4.5 12.C5.0 13. Oblique Classifier XGBoost	1. Support Vector Machine (SVM) 2. Neural Network	1. Decision Tree (not mentioned Decision Tree algorithm) 2. Gradient boosted tree 3. Gaussian Processes 4. Random Tree (GPRT) 5. Classification using CHAID (Chi-Squared Automatic Interaction Detection) 6. Discriminant analysis 7.GLM: Generalized Linear Model 8. Singular Value 9. Decomposition (SVD) Interpretable 10. Classification Rule Mining 11. MaxMargin Multi-Label (M3L) classifier 12. LIBSVM 13. Cross-out classifier 14. Classifier based on polynomial regression and stochastic gradient descent Improved ID3

Some of the important deep learning classification algorithms are given here. Artificial neural network (ANN) is a model consisting of several processing elements. These processing elements receive inputs and based on predefined activation functions, provides outputs. ANN has three or more layers such as input neurons, neural layers and last output layer.

A multilayer perceptron (MLP) is a class of artificial neural network (ANN). The MLP is used in stock market analysis, image identification, spam detection, etc.

Table 6 shows the comparative analysis based on classification techniques including reference number, publication year, article from which research database, performance parameters, dataset used, sampling period, tool/ software used and data collection technique. Performance evaluation measures considered are generally accuracy, precision, recall, f-measures, root mean square error, k-fold value, true positive rate, false positive rate, receiver operating characteristic curve, area under ROC curve, etc. Dataset used is generally students data, log files of students, discussion posts, programming language performance data, web based course performance data, etc. Sampling period considered is year or semester data. Data Mining tools applied to dataset are RapidMiner, SPSS, Weka, Orange, KNIME, etc while software used to build the model in EDM are Java, Python, R-programming, SQL, ASP.NET, ANSI C, MATLAB, etc. Data collection techniques considered for dataset from educational setting are log files from web based course, data related to threads and posts for course, survey data, course data Learning Management System/ Course Management System, MOODLE logs, etc. From Table 5, it is found that mostly used performance parameters for classification algorithms are accuracy, recall, precision, F-measure and k-fold value. It is also noted that Tools used are Weka, and RapidMiner while software used are Java, R-Programming, and Python. Mostly used classification algorithms are Naïve Bays (El-Halees, 2011; Chau and Phung, 2013; Pratiwi, 2013; Göker et al., 2013; Mashiloane and Mchunu, 2013; Palazuelos et al., 2013; Dangi and Srivastava, 2014; Anh et al., 2014; Chen et al., 2014; Ragab et al., 2014; Manhães et al., 2014; Pruthi and Bhatia, 2015; Guo et al., 2015; Guarín et al., 2015; Ahadi et al., 2015; Bakaric et al., 2015; Barbosa Manhães et al., 2015; Jishan et al., 2015; Salinas and Stephens, 2015; Kaur et al., 2015; Mayilvaganan and Kalpanadevi, 2015; Amornsinlaphachai, 2016; Devasia et al., 2016; Lehr et al., 2016; Chaudhury et al., 2016; Ahmed et al., 2016; Athani et al., 2017; Castro-Wunsch et al., 2017; Daud et al., 2017; Pise and Kulkarni, 2017; Costa et al., 2017; Kiu, 2018; Srivastava et al., 2018; Chitra and Agrawal, 2019; Niu et al., 2018; Rustia et al., 2018; Abyaa et al., 2018; Chanlekha and Niramitranon, 2018; Pérez et al., 2018; Maitra et al., 2018; Tasnim et al., 2019; Al Breiki et al., 2019; Lagman et al., 2019; Umer et al., 2019; Amazona and Hernandez, 2019; Francis and Babu, 2019; Adekitan and Salau, 2020; Tarmizi et al., 2019; Rawat and Malhan, 2019; Dimić et al., 2019; Vila et al., 2018; Ab Ghani et al., 2018; Ibrahim et al., 2018; Adekitan and Salau, 2019; Lottering et al., 2020; Mengash, 2020; Santoso, 2019; Injadat et al., 2020a; Karthikeyan et al., 2020; Ajibade et al., 2018; El Aissaoui et al., 2020; Ashraf et al., 2020; Injadat et al., 2020b), Support Vector Machine (Cocea and Weibelzahl, 2010; El-Halees, 2011; Dejaeger et al., 2012; Chau and Phung, 2013; Anh et al., 2014; Chen et al., 2014; Manhães et al., 2014; Guo et al., 2015; Sorour et al., 2015; Bakaric et al., 2015; Barbosa Manhães et al., 2015; Agaoglu, 2016; Chaudhury et al., 2016; Buenaño-Fernández et al., 2017; Kitanaka et al., 2017; Daud et al., 2017; Leppänen et al., 2017; Costa et al., 2017; Srivastava et al., 2018; Chitra and Agrawal, 2019; Lagus et al., 2018; Spatiotis et al., 2018; Niu et al., 2018; Rustia et al., 2018; Abyaa et al., 2018; Chanlekha and Niramitranon, 2018; Tasnim et al., 2019; Francis and Babu, 2019; Tarmizi et al., 2019; Zaffar et al., 2018; Ibrahim et al., 2018; Lottering et al., 2020; M.Á. et al., 2020; Mengash, 2020; Rahman and Mahmud, 2020; Mkwazu and Yan, 2020; Injadat et al., 2020a; Agrawal et al., 2020; Ribeiro and Canedo, 2020; Injadat et al., 2020b), Logistic Regression (Cocea and Weibelzahl, 2010; Chuan et al., 2011; Chau and Phung, 2013; Barbosa Manhães et al., 2015; Lehr et al., 2016; Jung, 2016; Rustia et al., 2018; Abyaa et al., 2018; Pérez et al., 2018; Burgos et al., 2018; Tasnim et al., 2019; Alshehri,

Table 6
Analysis based on classification technique.

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Guruler et al. (2010)	Elsevier Journal	2010	A new student performance analyzing system using knowledge discovery in higher educational databases	Decision Tree	Lift value	Students data	–	SQL Server 2000	Student data from the directory of student affairs of the university	–
Cocea and Weibelzahl (2010)	IEEE Transac-tion	2011	Disengagement Detection in Online Learning: Validation Studies and Perspectives	Bayesian Nets, Logistic regression, Simple logistic classification, Instance-based classification with IBk algorithm, Attribute Selected Classification using J48 classifier and Best First search, Bagging using REP (Reduced error pruning) tree classifier, Classification via Regression, Decision Trees with J48 classifier based on Quilan's C4.5	Accuracy, True positive (TP) rate, False positive (FP) rate, Precision, Error	Log files of 48 users	–	–	Log files from web-based interactive learning environment	87%
Chuan et al. (2011)	Springer Proceed-ing	2011	Combining Different Classifiers in Educational Data Mining	Logistic Regression, K-nearest Neighbour (KNN) and Singular Value Decomposition (SVD)	Root mean square error (RMSE)	9353 students	2008–2009	–	Student data	–
El-Halees (2011)	Springer Proceed-ing	2011	Mining Opinions in User-Generated Contents to Improve Course Evaluation	K-nearest,Naïve Bays, Support Vector Machine	Precision, Recall, F-measures, Cross-validation K-fold	4,957 discussion posts	–	RapidMiner	Five courses content including all threads and posts about courses	–
Wang and Liao (2011)	Elsevier Journal	2011	Data mining for adaptive learning in a TESL-based e-learning system	Artificial Neural Network	Root mean square error (RMSE)	A total of 70 freshmen at a university in central Taiwan	–	–	Students data	–
Sakurai et al. (2012)	IEEE Pro-ceeding	2012	A Case Study on Using Data Mining for University Curricula	Decision Tree	Cosine Cofficient	Student data	–	–	Former students curricula and their degree of success	–
Bunkar et al. (2012)	IEEE Pro-ceeding	2012	Data Mining: Prediction for Performance Improvement of Graduate Students using Classification	ID3 (Iterative Dichotomiser 3), C4.5 and CART (Classification and Regression Trees)	True positive rate, False positive rate, Precision, Recall, Receiver Operating Characteristic, F-measures	BA First Year student data with 6 attributes (Year 2009)	–	–	BA First Year Student data	–
Nasiri et al. (2012)	IEEE Pro-ceeding	2012	Predicting GPA and Academic Dismissal in LMS Using Educational Data Mining: A Case Mining	Regression and C5.0	Accuracy	Student data with 13 fields	–	SPSS Clementine 12.0	Real data of an e-Learning system	88.50%
Dejaeger et al. (2012)	Elsevier Journal	2012	Gaining insight into student satisfaction using comprehensible data mining techniques	Classification And Regression Trees (CART), Multilayered perceptrons (MLP), Oblique Classifier 1 (OC1), Logit, Support Vector Machine (SVM)	Accuracy, Area under Curve (AUC)	8550 Students with 15 question attributes	2007–2008.	–	Survey	75%

(*continued on next page*)

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Şen et al. (2012)	Elsevier Journal	2012	Predicting and analyzing secondary education placement-test scores: A data mining approach	C5 Decision tree	Accuracy, Cross-validation K-fold	5000 unique student records	2009	–	Student data	95%
Sen and Ucar (2012)	Elsevier Proceeding	2012	Evaluating the achievements of computer engineering department of distance education students with data mining methods	Artificial Neural Networks, and Decision trees	Accuracy, Cross-validation K-fold	3047 records with 5 attributes	–	–	Student data	97.81%
Abaya and Gerardo (2013)	IEEE Proceeding	2013	An Education Data Mining Tool for Marketing based on C4.5 Classification Technique	C4.5	–	1497 instances. n the historical data	–	–	Historical data of College Entrance Test examinees	–
Hoe et al. (2013)	IEEE Proceeding	2013	Analyzing Students Records to Identify Patterns of Students' Performance	Classification using CHAID (Chi-Squared Automatic Interaction Detection)	Accuracy	2,228 records of the Foundation students with 8 attributes	2004–2013	–	Student data	70.17%
Chau and Phung (2013)	IEEE Proceeding	2013	Imbalanced educational data classification: an effective approach with resampling and random forest	Naïve Bayes, Support Vector Machine (SVM), Neural Network, Logistic Regression, K-nearest Neighbour (KNN), and Decision Ttree C4.5, Random Forest	Accuracy, Receiver Operating Characteristic, Cross-validation K-fold	5336 student data with 43 attributes	2005–2008	Weka tool	Student data	94.15%
Pratiwi (2013)	IEEE Proceeding	2013	Predicting Student Placement Class using Data Mining	J48, SimpleCart, Kstar, SMO, Naive Bayes, OneR	Accuracy, Correctly Classified Instances, Incorrectly Classified Instances, Cross-validation K-fold	314 students with 33 attributes	2010–2011	–	Scores of students in 23 senior high school	79.61%
Göker et al. (2013)	IEEE Proceeding	2013	The Estimation of Students' Academic Success by Data Mining Methods	Naive Bayes, J48, Bays Net, RBF Network	Accuracy, Precision, Recall, Receiver Operating Characteristic, F-measures, Correctly Classified Instances, Incorrectly Classified Instances	200 with 34 attributes, student records	–	Weka	Student information of the high school	85%
Márquez-Vera et al. (2013b)	IEEE Transaction	2013	Predicting School Failure and Dropout by Using Data Mining Techniques	Jrip, Nnge, OneR, Prism, Ridor, J48, C4.5, SimpleCart, ADTree, Random Tree	Accuracy, True positive (TP) rate, True Negatives (TN) rate, Cross-validation K-fold	670 middle-school students with 77 attributes	2009–2010	Weka	Student data	97%
Márquez-Vera et al. (2013a)	Springer Journal	2013	Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data	rip, Nnge, OneR, Prism, Ridor, J48, C4.5, SimpleCart, ADTree, Random Tree, REPTree, Interpretable Classification Rule Mining(ICRM)	Accuracy, True positive (TP) rate, True Negatives (TN) rate, Cross-validation K-fold	670 high school students with 77 attributes from Zacatecas, Mexico	August - December 2010	Weka	Student data	93%

(*continued on next page*)

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Mashiloane and Mchunu (2013)	Springer Book	2013	Mining for Marks: A Comparison of Classification Algorithms when Predicting Academic Performance to Identify "Students at Risk"	J48 classifier, Naïve Bayes and Decision Table	Accuracy, False positive rate, Precision, Correctly Classified Instances, Incorrectly Classified Instances	391 students with 8 attributes registered for the CS-1 unit	2009–2011	Weka	Student data	87.55%
Palazuelos et al. (2013)	Springer Proceed-ing	2013	Social Network Analysis and Data Mining: An Application to the E-Learning Context	J48, random forests, Naïve Bayes, Bayesian networks, JRip, and Ridor	Accuracy, Cross-validation K-fold	194 instances at the UC.	2007–2008, 2008–2009, and 2009–2010	Weka	Introduction to Multimedia Methods" Course data hosted on Black-board/WebCT, entitled	70%
Blagojević and Micić (2013)	Elsevier Journal	2013	A web-based intelligent report e-learning system using data mining techniques	Decision Tree and Neural Network	Probability	The number of active users, according to the system logs, amounted to 1935	–	SQL Server	Moodle Log	–
Dangi and Srivastava (2014)	IEEE Pro-ceeding	2014	Educational data Classification using Selective Naïve Bayes for Quota categorization	Naïve Bayes	Correctly Classified Instances, Incorrectly Classified Instances, Mean Absolute Error, Root mean square error, Relative absolute error, Root relative squared error	152 entries with 42 attributes from the Center of Converg-ing Technolo-gies	–	–	Student data	–
Pathan et al. (2014)	IEEE Pro-ceeding	2014	Educational Data Mining: A Mining Model for Developing Students' Programming Skills	ID3 and C4.5	Accuracy	70 students with 16 attributes of Structured Program-ming Language course	–	Weka	Student data	87%
Anh et al. (2014)	IEEE Pro-ceeding	2014	Towards a robust incomplete data handling approach to effective educational data classification in an academic credit system	Naïve Bayes, Neural Network, Support Vector Machine, K-Nearest Neighbour (K-nn), Decision Tree C4.5, and Random Forest	Accuracy, Receiver Operating Characteristic (ROC), Cross-validation K-fold	1334 students with 43 attributes , grade informa-tion of under-graduate students enrolled in 2005–2008	–	–	Student data	–

(*continued on next page*)

2019; Umer et al., 2019; Al Fanah and Ansari, 2019; Almutairi et al., 2019; Ab Ghani et al., 2018; Adekitan and Salau, 2019; Mkwazu and Yan, 2020; Islam and Mahmud, 2020; Injadat et al., 2020a,b), K-nearest neighbour (Chuan et al., 2011; El-Halees, 2011; Chau and Phung, 2013; Anh et al., 2014; Amornsinlaphachai, 2016; Lehr et al., 2016; Pise and Kulkarni, 2017; Srivastava et al., 2018; Niu et al., 2018; Abyaa et al., 2018; Al Breiki et al., 2019; Rawat and Malhan, 2019; Yousafzai et al., 2020; Lottering et al., 2020; Rahman and Mahmud, 2020; Islam and Mahmud, 2020; Injadat et al., 2020a; Ajibade et al., 2018; Agrawal et al., 2020; Ashraf et al., 2020; Injadat et al., 2020b), J48 (Bodea et al., 2010; Chellatamilan et al., 2011; Cocea and Weibelzahl, 2010; Trandafili et al., 2012; Pratiwi, 2013; Göker et al., 2013; Márquez-Vera et al., 2013b,a; Mashiloane and Mchunu, 2013; Palazuelos et al., 2013; Natek and Zwilling, 2014; Parmar et al., 2015; Ahadi et al.,

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Chen et al. (2014)	IEEE Transaction	2014	Mining Social Media Data for Understanding Students' Learning Experiences	Naïve Bayes , MaxMargin Multi-Label (M3L) classifier, SVM	Accuracy, Precision, Recall, F-measures	samples taken from about 25,000 tweets related to engineering students' college life	Nov. 1st, 2011 - Dec. 25th, 2012	–	Students' Twitter posts to understand issues and problems in their educational experiences.	–
Ragab et al. (2014)	ACM Proceeding	2014	A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining	C4.5, Random Forest, IBK-E, IBK-M, LibSVM, Multilayer Perceptron, Naïve Bayes, and PART	True positive (TP) rate, False positive (FP) rate,Precision, Recall, Receiver Operating Characteristic (ROC), F-measures, Cross-validation K-fold	5260 instances with 8 attributes	–	Weka	Database of the preparatory year students	–
Manhães et al. (2014)	ACM Proceeding	2014	WAVE: an Architecture for Predicting Dropout in Undergraduate Courses using EDM	Naïve Bayes (NB), Multilayer Perceptron (MLP),Support Vector Machine with polynomial kernel (SVM1) and RBF kernel (SVM2) and Decision Table	Accuracy, True positive (TP) rate, False positive (FP) rate, False Negatives (FN) Rate, True Negatives (TN) rate, Cross-validation K-fold	Student Data	1994–2010.	–	Student Data	91.34%
Natek and Zwilling (2014)	Elsevier Journal	2014	Student data mining solution–knowledge management system related to higher education institutions	J48, M5P and RepTree	Accuracy	106 students with 12 attributes	2010–2011, 2011–2012, 2012–2013	Weka	Students data	97%
Gera and Goel (2015)	IEEE Proceeding	2015	A Model for Predicting the Eligibility for Placement of Students Using Data Mining Technique	Decision Tree	True positive rate, False positive rate, Precision, Recall, Receiver Operating Characteristic, F-measures, Mean Absolute Error, Root mean square error, Relative absolute error, Root relative squared error	Student dataset	–	ASP.NET as front end and SQL for database	Student dataset of University placement centre	–
Pruthi and Bhatia (2015)	IEEE Proceeding	2015	Application of Data Mining in Predicting Placement of Students	148 decision tree algorithm,NaiVe Bayes	Accuracy, Mean Absolute Error(MAE)	424 instances with 26 attributes in the dataset	–	Weka	Student data	62.10%
Parmar et al. (2015)	IEEE Proceeding	2015	Performance prediction of students using distributed data mining	Random Forest and J48	Accuracy	Parul institute students' dataset of each having 1000 records.	–	Weka, Java	Student data	–
Guo et al. (2015)	IEEE Proceeding	2015	Predicting Students Performance in Educational Data Mining	Naïve Bayes, Multilayer Perception (MLP) and SVM ,SPPN Prediction Network	Precision	1200 grade-9 students for recent three years	–	ANSI C and Theano, a python library that allows transparent use of GPU	Data from 100 junior high schools	–

(*continued on next page*)

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Guarín et al. (2015)	IEEE Transaction	2015	A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining	Naïve Bayes and Decision Tree classifier	Accuracy, Cross-validation K-fold	1532 records.	The second academic period of 2007 and the second academic period of 2012	–	Student data from two programs, Agricultural (AE) and Computer and Systems (CE) Engineering,	54%–57%
Sorour et al. (2015)	ACM Proceeding	2015	Correlation of Topic Model and Student Grades Using Comment Data Mining	Support Vector Machine (SVM) and Artificial Neural Network (ANN)	Accuracy, Precision, Recall, F-measures, Cross-validation K-fold	Comment data from 123 students	–	MATLAB LibSVM tool	Comment data	74%
Ahadi et al. (2015)	ACM Proceeding	2015	Exploring Machine Learning Methods to Automatically Identify Students in Need of Assistance	Naive Bayes, Bayesian Network, Decision Table, Conjuctive Rule, PART, ADTree, J48, Random Forest, Decision Stump	Accuracy, Cross-validation K-fold	In the first semester (spring), a total of 86 students partici-pated in the study, and in the second semester (fall), a total of 210 students partici-pated in the study.	–	–	Student data	74.66%
Sisovic et al. (2015)	ACM Proceeding	2015	Mining Student Data to Assess the Impact of Moodle Activities and Prior Knowledge on Programming Course Success	C4.5, JRip and PART	Accuracy, Cross-validation K-fold	153 instances of the course Program-ming 1	Two winter semesters -2013–2014 and 2014–2015	Weka	the first part is extracted from the Learning Management System Moodle logs, while the second part is related to prior knowledge and students' preparation for the study.	83%
Bakaric et al. (2015)	ACM Proceeding	2015	Text Mining Student Reports	C4.5, PART, NaiveBayes, and SVM	Accuracy, Cross-validation K-fold	52 Student data	–	–	Student data	83.50%
Bar-bosa Man-hães et al. (2015)	ACM Proceeding	2015	Towards Automatic Prediction of Student Performance in STEM Undergraduate Degree Programs	BayesNet (BN), NaiveBayes (NB), J48, JRip (JR), Support Vector Machines, Multilayered Perceptrons (MLP), AdaBoost (AB), SimpleLogistic (SL), Decision Table (DT), OneR, and Random Forest (RF).	True positive rate, False positive rate, Precision, Recall, Confusion Matrix,Correctly Classified Instances , Incorrectly Classified Instances	402 students	–	Weka	The dataset provided by UFRJ's academic registry	–

(*continued on next page*)

2015; Sisovic et al., 2015; Barbosa Manhães et al., 2015; Auddy and Mukhopadhyay, 2015; Kaur et al., 2015; Sanchez-Santillan et al., 2016; Meedech et al., 2016; Ahmed et al., 2016; Angra and Ahuja, 2017; Ayub et al., 2017; Castro-Wunsch et al., 2017; Pise and Kulkarni, 2017; Kiu, 2018; Chitra and Agrawal, 2019; Spatiotis et al., 2018; Abyaa et al., 2018; Martínez-Abad et al., 2018; Kularbphettong, 2017; Chango et al., 2019; Tarmizi et al., 2019; Rawat and Malhan, 2019; Dimić et al., 2019; Karthikeyan et al., 2020; Ashraf et al., 2020),

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Jishan et al. (2015)	Springer Book	2015	Application of Optimum Binning Technique in Data Mining Approaches to Predict Students' Final Grade in a Course	Naive Bayes, C4.5, Neural Network	Accuracy, Precision, Recall, Area under Curve, F-measures	181 instances	Record of five semesters	RapidMiner	Course grade data	68.89%
Salinas and Stephens (2015)	Springer Proceeding	2015	Applying Data Mining Techniques to Identify Success Factors in Students Enrolled in Distance Learning: A Case Study	Naive Bayes	True positive rate, Area under Curve	2,889 students with 29 attributes	2005–2010	–	Students' data from the Open University System and Distance Learning	–
Auddy and Mukhopadhyay (2015)	Springer Proceeding	2015	Data Mining on ICT Usage in an Academic Campus: A Case Study	J48, JRip, QuickRules Fuzzy-Rough rule induction, Fuzzy Nearest Neighbour, Fuzzy Rough Nearest Neighbour and Vaguely Quantified Nearest Neighbour (VQNN)	True positive rate, False positive rate, Precision, Recall, Receiver Operating Characteristic, F-measures, Cross-validation K-fold	Student data with 15 attributes	–	Weka	Responses collected from the student and research communities via survey on ICT	–
Shukor et al. (2015)	Elsevier Proceeding	2015	An examination of online learning effectiveness using data mining	C4.5	t-Test	20 undergraduate student data about Web-based course Multimedia Development	–	Weka	Student data from online learning course	–
Kaur et al. (2015)	Elsevier Proceeding	2015	Classification and prediction based data mining algorithms to predict slow learners in education sector	Multilayer Perception, Naïve Bayes, SMO, J48 and REPTree	Accuracy, True positive rate, False positive rate, Precision, Recall, Receiver Operating Characteristic, F-measures	152 students with 14 attributes of high school	–	Weka	Students data	75%
Amornsinlaphachai (2015)	Elsevier Proceeding	2015	The design of a framework for cooperative learning through web utilizing data mining technique to group learners	ID3	–	–	–	–	–	–
Amornsinlaphachai (2016)	IEEE Proceeding	2016	Efficiency of data mining models to predict academic performance and a cooperative learning model	Artificial Neural Network, K-Nearest Neighbour, Naive Bayes, Bayesian Belief Network, JRIP, ID3 and C4.5	Precision, Recall, F-measures, Cross-validation K-fold, Mean Absolute Error(MAE)	474 students with 12 attributes	–	–	Student Data	–

(*continued on next page*)

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Devasia et al. (2016)	IEEE Proceeding	2016	Prediction of Students Performance using Educational Data Mining	Naive Bayesian		700 students' with 19 attributes	2013–2016	Weka	Student Data	–
Lehr et al. (2016)	IEEE Proceeding	2016	Use Educational Data Mining to Predict Undergraduate Retention	Logistic Regression, Naïve Bayes, K Nearest Neighbourhood (KNN), Random Forest, Multilayer Perceptron (MLP), and Decision Tree	Receiver Operating Characteristic, Cross-validation K-fold	972 students enroll-ment data with 38 attributes	2008	Weka	Student Data	–
Agaoglu (2016)	IEEE Transaction	2016	Predicting Instructor Performance Using Data Mining Techniques in Higher Education	C5.0, CART , Support Vector Machines, Artificial Neural Networks, and Discriminant Analysis	Accuracy, Precision, Recall	2850 evaluation scores	–	–	Student data	92.30%
Chaud-hury et al. (2016)	ACM Proceeding	2016	Enhancing the capabilities of Student Result Prediction System	SVM, C4.5 and Naïve Bayes	True positive rate, False positive rate, Precision, Receiver Operating Characteristic	33 attributes with 678 instances.	–	Weka	The dataset from the UCI repository	–
Sanchez-Santillan et al. (2016)	ACM Proceeding	2016	Predicting Students' Performance: Incremental Interaction Classifiers	JRIP (a RIPPER implementation) , J48 and Bayesian Network	Accuracy, Cross-validation K-fold	195 students' interac-tion in two different years;	2012 (N = 111 students) and 2013 (N= 84 students)	Weka	Student data	76%
Ra-manathan et al. (2016)	Springer Book	2016	Apply of Sum of Difference Method to Predict Placement of Students' using Educational Data Mining	Decision Tree and Random Forest	Sum of Difference (SOD)	50 pupils(s) with 11 given character-istics	–	C# programming language and MS Visual Studio 2012	Student data	–
Meedech et al. (2016)	Springer Book	2016	Prediction of Student Dropout using Personal Profile and Data Mining Approach	JRip, OneR, Ridor, J48, SimpleCart, ADTree, RandomTree and REPTree	Accuracy, Cross-validation K-fold, Standard Deviation	509 records with 19 attributes	2012–2013	Weka	Data of students at Mae Fah Luang University.	80%
Stahovich and Lin (2016)	Elsevier Journal	2016	Enabling data mining of handwritten coursework	Cross-out classifier	Accuracy, Cross-validation K-fold	60,000 pages of ink with 28 106 time-stamped pen strokes from 700 students	–	–	Student data	90%
Hassan and Al-Razgan (2016)	Elsevier Proceeding	2016	Pre-University Exams Effect on Students GPA: A case Study in IT Department	Linear Regression	Standard Deviation	957 students data	–	Weka	Student data from the ad-ministration.	67.33%

(*continued on next page*)

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Hamsa et al. (2016)	Elsevier Proceeding	2016	Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm	Decision tree - C4.5	Splitting Attribute	168 students data	–	–	Students data	–
Ahmed et al. (2016)	Elsevier Proceeding	2016	Using data mining to predict instructor performance	J48 , Multilayer Perceptron, Naïve Bayes , Sequential Minimal Optimization	Accuracy, Cross-validation K-fold	5820 evaluation scores provided by students	–	Weka	University of California-Irvine (UCI) Machine Learning Repository	85%
Kassak et al. (2016)	Elsevier Journal	2016	Student behavior in a web-based educational system:Exit intentprediction	Classifier based on Polynomial Regression and Stochastic Gradient Descent	Accuracy, Precision	160 bachelor students of Software engineering course	–	–	e-learning students data	80%
Athani et al. (2017)	IEEE Proceeding	2017	Student Academic Performance and Social Behavior Predictor using Data Mining Techniques	Naïve Bayes	Accuracy, Cross-validation K-fold	395 tuples	–	Weka	Real data was collected using school reports and questionnaire method	87%
Castro-Wunsch et al. (2017)	ACM Proceeding	2017	Evaluating Neural Networks as a Method for Identifying Students in Need of Assistance	NaiveBayes, Neural Network, RandomForest, DecisionTable, DecisionStump, PART, J48	Accuracy	1160 students	Fall 2014, Fall 2015 and Winter 2016	Weka	Student course data	70.43%
Buenaño-Fernández et al. (2017)	ACM Proceeding	2017	Improvement of massive open online courses by text mining of students' emails: a case study	Support Vector Machine (SVM)	Accuracy	Number of documents 950 Number of tokens 82,174 Number of word types 7,890	–	–	The emails data	75%
Figueira (2017)	ACM Proceeding	2017	Mining Moodle Logs for Grade Prediction: A methodology walk-through	Random Forest	Accuracy	Students: 161 Resources: 416 Events: 64	–	–	Moodle LMS during the period of one academic semester	99%
Kitanaka et al. (2017)	ACM Proceeding	2017	Predicting Learning Result of Learner in E-learning Course with Feature Selection Using SVM	Support Vector Machine (SVM)	Accuracy, Precision, Recall, F-measures	Student data	–	–	Learning contents from the Virtual Learning Environment (VLE)	77%
Daud et al. (2017)	ACM Proceeding	2017	Predicting Student Performance using Advanced Learning Analytics	Support Vector Machine (SVM), C4.5, Classification and Regression Tree (CART), Bayes Network (BN), Naive Bayes (NB)	F-measures, Cross-validation K-fold	776 student instances for experiments.	2004–2011	Weka	Graduate and undergraduate students data	–
Leppänen et al. (2017)	ACM Proceeding	2017	Predicting Academic Success Based on Learning Material Usage	Support Vector Methods: a linear kernel, an RBF kernel and a sigmoid kernel, Regression analysis	Accuracy, Cross-validation K-fold	271 participants participated in the final exam	Fall 2016	–	A seven week - Introduction to Programming course	–

(*continued on next page*)

and Random Forests (Chau and Phung, 2013; Márquez-Vera et al., 2013b; Palazuelos et al., 2013; Anh et al., 2014; Ragab et al., 2014; Parmar et al., 2015; Ahadi et al., 2015; Barbosa Manhães et al., 2015; Lehr et al., 2016; Meedech et al., 2016; Castro-Wunsch et al., 2017; Figueira, 2017; Pise and Kulkarni, 2017; Kiu, 2018; Kaunang and Rotikan, 2018; Chitra and Agrawal, 2019; Lagus et al., 2018; Spatiotis et al., 2018; Abyaa et al., 2018; Chanlekha and Niramitranon, 2018; Pérez et al., 2018; Miguéis et al., 2018; Al Breiki et al., 2019;

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Pise and Kulkarni (2017)	Springer Journal	2017	Evolving learners' behavior in data mining	K-nearest Neighbour	Accuracy, Cross-validation K-fold	38 benchmark data sets from the University of California at Irvine Machine Learning Repository	–	Weka	–	–
Jung (2016)	Springer Proceeding	2017	A Comparison of Data Mining Methods in Analyzing Educational Data	Neural Network - Multi-Layer Perceptron (MLP), Logistic Regression, Decision Tree -Chi-square Automatic Interaction Detector (CHAID) for Decision Tree	Accuracy	3,449 data	2003–2008	–	Student data	83.60%
Costa et al. (2017)	Elsevier Journal	2017	Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses	Neural Networks, Decision Tree, Support Vector Machine, Naive Bayes	F-measures	1, 262 undergraduate students performed	2014	Java	Distance education modality and On-campus	–
Kiu (2018)	IEEE Proceeding	2018	Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities	Naïve Bayesian, Multilayer Perceptron, Decision Tree J48 and Random Forest	Precision, Recall, F-measures	395 instances with 33 attributes	–	–	Students' performance in Mathematics subjects	–
Srivastava et al. (2018)	IEEE Proceeding	2018	Educational Data Mining: Classifier Comparison for the Course Selection Process	K-NN, Support Vector machine and Naïve Bays	Accuracy	2890 records entries with 15 attributes	–	–	Engineering student records for the open elective choices of interdisciplinary domain.	98%
Patil et al. (2018)	IEEE Proceeding	2018	Prediction System for Student Performance Using Data Mining Classification	ID3, improved ID3 and C4.5	Accuracy	Student data	–	–	Training database	74%
Kaunang and Rotikan (2018)	IEEE Proceeding	2018	Students' Academic Performance Prediction using Data Mining	Random Forest and Decision Tree	Accuracy, Precision, Recall, F-measures, Cross-validation K-fold	249 records with 3 different classes.	–	Weka	Questionnaire consisted of 7 questions	64%
Chitra and Agrawal (2019)	IEEE Proceeding	2019	Analysis of Educational Data Mining using Classification	J48, Support Vector machine, Naïve Bayes, Random Forest, Multilayer Perceptron	Accuracy	480 student records with 16 attributes	–	Weka	Collection of data from learner activity tracker tool	76.07%
Lagus et al. (2018)	ACM Transaction	2018	Transfer-Learning Methods in Programming Course Outcome Prediction	Support Vector Machine, Random Forest, and AdaBoost	Precision, Recall, F-measures, Cross-validation K-fold	348 student data	–	–	The data from two introductory Java programming courses	–

(*continued on next page*)

Umer et al., 2019; Al Fanah and Ansari, 2019; Adekitan and Salau, 2020; Tarmizi et al., 2019; Martins et al., 2019; Dimić et al., 2019; Almutairi et al., 2019; Kasthuriarachchi and Liyanage, 2018; Ibrahim et al., 2018; Adekitan and Salau, 2019; Utari et al., 2020; Injadat et al., 2020a; Agrawal et al., 2020; Ribeiro and Canedo, 2020; Injadat et al., 2020b).

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Spatiotis et al. (2018)	ACM Proceeding	2018	Evaluation of an Educational Training Platform Using Text Mining	REPTree, CART, Random Forest, J48, Ibk, Bagging, AdaBoostM1, SVM, Neural Network	Accuracy, Cross-validation K-fold	2600 teacher data	–	–	set of opinions collected from the questionnaires in the e-learning courses through HEP system	63%
Niu et al. (2018)	ACM Proceeding	2018	Exploring Causes for the Dropout on Massive Open Online Courses	Generalized Linear Models, Support Vector Machines, Tree Based Ensemble Learning Models, Neural Network Models, Naïve Bayes Models, K-Neighbours, Gaussian Progress	Accuracy, Area under Curve, Cross-validation K-fold	More than 1,600,000 click records and 150,000 table records	–	–	Student data from 163 platform	94%
Rustia et al. (2018)	ACM Proceeding	2018	Predicting Student's Board Examination Performance using Classification Algorithms	Naïve Bayes, Support Vector Machine, Neural Network, C4.5 Decision Tree, Logistic Regression	Accuracy, Area under Curve , Cross-validation K-fold	446 students	2013–2016	RapidMiner	The dataset comes from student academic record	73%
Abyaa et al. (2018)	ACM Proceeding	2018	Predicting the learner's personality from educational data using supervised learning	Support Vector Machines (SVM), k-Nearest Neighbours (kNN), Naïve Bayes, Random forest, J48, Logistic regression, Bagging	Correctly Classified Instances, Cross-validation K-fold, Mean Absolute Error	48 students over a 10-week term	–	Weka	Student data	–
Chanlekha and Niramitra-non (2018)	ACM Proceeding	2018	Student Performance Prediction Model for Early-Identification of At-risk Students in Traditional Classroom Settings	Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine, Artificial Neural Network	Accuracy, Precision, Recall, F-measures, Cross-validation K-fold	Student grade and demographic information	2008–2017	–	Student data	–
Martínez-Abad et al. (2018)	ACM Proceeding	2018	Big Data in Education: Detection of ICT Factors Associated with School Effectiveness with Data Mining Techniques	J48	True positive rate, Precision, Receiver Operating characteristic, Kappa Statistic, Cross-validation K-fold, Root relative squared error	130 selected as having high effectiveness and 127 of low effectiveness.	–	Weka	Student data	–
Ku-larbphet-tong (2017)	Springer Proceeding	2018	Analysis of Students' Behavior Based on Educational Data Mining	J48, Bayesian Networks	Accuracy, Precision, Recall, F-measures	5392 student personal records	2014–2015	Weka	Student personal records	91%
Pérez et al. (2018)	Springer Proceeding	2018	Predicting Student Drop-Out Rates Using Data Mining Techniques: A Case Study	Decision Trees, Logistic Regression, Naive Bayes and Random Forest	Area under Curve (AUC), Cross-validation K-fold	Student data	2004–2010	Watson Analytics, Python	Systems Engineering (SE) undergraduate student data	–

(*continued on next page*)

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Burgos et al. (2018)	Elsevier Journal	2018	Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout	Logistic Regression	Accuracy, Precision, Recall, Cross-validation K-fold	104 students	–	–	students data generated in Moodle	97%
Miguéis et al. (2018)	Elsevier Journal	2018	Early segmentation of students according to their academic performance: A predictive modeling approach	Random Forest	Accuracy, Cross-validation K-fold	2459 student data	2003–2015	–	Students data	96.10%
Maitra et al. (2018)	Elsevier Proceeding	2018	Mining authentic student feedback for faculty using Naïve Bayes classifier	Naíve Bayes	–	1000 students of an Indian Higher Education Institution affiliated with a state university	–	R-Programming	Students data	–
Tasnim et al. (2019)	IEEE Proceeding	2019	Identification of DropOut Students Using Educational Data Mining	Logistic Regression, Naive Bayes Classifier, Support Vector Machine and Threshold based approach for classification	Precision, Recall, F-measures, Cross-validation K-fold	395 instances and 649 instances with 33 attributes in each	–	MATLAB R2017a.	Student data	–
Ketui et al. (2019)	IEEE Proceeding	2019	Using Classification Data Mining Techniques for Students Performance Prediction	Decision Tree, Decision Tree Weight-based, ID3, Random Tree, and gradient boosted tree (GBT)	Accuracy, Cross-validation K-fold	17,875 academic achievements within 483 students.	2009–2013	–	Students of Faculty of Science and Agricultural Technology	92%
Al Breiki et al. (2019)	IEEE Proceeding	2019	Using Educational Data Mining Techniques to Predict Student Performance	Simple Log Regress (SLR), Decision Table (DT), Gaussian Processes Random Tree (GPRT), propositional rule learner (JRip), K-nearest neighbours (IBK), random forest, multi-layer perceptron, Naïve Bayes, Bayes Network (BN) learning, Multilayer Perceptron (MP), Linear Regression (LR), Random Forest (RF), and Sequential Minimal Optimization	Accuracy, Correctly Classified Instances, Cross-validation K-fold, Mean Absolute Error(MAE), Root mean square error (RMSE), Relative absolute error (RAE), Root relative squared error (RRSE)	145 instances	Academic years 2009–2010 and 2014–2015.'-	Weka	Dataset from Student Information System	84%
Alshehri (2019)	ACM Proceeding	2019	Applying explanatory analysis in education using different regression methods	Logistic Regression	Precision, Recall, F-measures, Cross-validation K-fold	Student data	–	–	Student data	–
Lagman et al. (2019)	ACM Proceeding	2019	Embedding Naïve Bayes Algorithm Data Model in Predicting Student Graduation	Naíve Bayes	Accuracy	1164 students' academic data	–	Weka	Student data	85.22%
Umer et al. (2019)	ACM Proceeding	2019	Mining Activity Log Data to Predict Student's Outcome in a Course	Random Forest classifier (RF), Naïve Bayes (NB), Logistic regression (LR), Linear Discriminating analysis (LDA)	Cross-validation K-fold, Standard Deviation	400 records of students	–	Python	Assessment scores extracted from student management system (SMS) and weekly activity log from learning management system (LMS)	–

(*continued on next page*)

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Amazona and Hernandez (2019)	ACM Proceeding	2019	Modeling Student Performance Using Data Mining Techniques: Inputs for Academic Program Development	Naïve Bayes, Deep Learning in Neural Network, and Decision Tree	Accuracy, Cross-validation K-fold	300 students' intake with 9 attributes	2015–2018.	RapidMiner	Student data	95%
Chango et al. (2019)	ACM Proceeding	2019	Predicting academic performance of university students from multi-sources data in blended learning	J48, REPTREE, RamdomTree, JRIP, Nnge, PART	Receiver Operating Characteristic, F-measures, Correctly Classified Instances, Cross-validation K-fold	65 university student data	–	Weka	Student data	–
Altaf et al. (2019)	ACM Proceeding	2019	Student Performance Prediction using Multi-Layers Artificial Neural Networks: A Case Study on Educational Data Mining	Multi-Layer Perceptions	Accuracy, Recall	Log information about 10 courses with a total of 900 students.	2016–2017	–	Log file of their Campus Management System (CMS)	84.29%
Al Fanah and Ansari (2019)	ACM Proceeding	2019	Understanding E-learners' Behavior Using Data Mining Techniques	Random Forests, Logistic Regressions and Bayesian Networks	Accuracy, Cross-validation K-fold	Student data	–	Weka, Spyder-Python	Kaggle website	80%
Adekitan and Salau (2020)	Springer Journal	2019	Towards an improved learning process: the relevance of ethnicity to data mining prediction of students' performance	Tree, Naive Bayes, Random forest, Neural network	Accuracy	Statistical attributes of 2413 student	–	Orange	Dataset across the 4 colleges	79.80%
Tarmizi et al. (2019)	Springer Proceeding	2019	A Case Study on Student Attrition Prediction in Higher Education Using Data Mining Techniques	J48, Random Forest, Naïve Bayes (NB), SVM(RBF Kernal), SVM(Polynomial Kernal)	Accuracy	74, 669 instances with 31 attributes related to graduate information	2013–2018	Weka	Data by Center for Strategic Planning and Information	97.52%
Martins et al. (2019)	Springer Proceeding	2019	A Data Mining Approach for Predicting Academic Success – A Case Study	Random Forest	Cross-validation K-fold, Root mean square error (RMSE)	4530 matriculations with 45 attributes	2007–2008 and 2015–2016 .	–	Studenr data	–
Rawat and Malhan (2019)	Springer Proceeding	2019	A Hybrid Classification Method Based on Machine Learning Classifiers to Predict Performance in Educational Data Mining	J48, K-nearest Neighbour, Naïve Bayes, Multilayer Perception	Accuracy, Cross-validation K-fold	27 instances and 11 attributes.	–	Weka	Student data	93%

(*continued on next page*)

3.3. Classification and clustering technique

This section illustrates the classification and clustering technique in EDM for determining students' performance. The research article Bodea et al. (2010), Chellatamilan et al. (2011), Bresfelean et al. (2012), Akram et al. (2019), Francis and Babu (2019), M.Á. et al. (2020), and Santoso (2019) applied the k-means clustering algorithm on the educational dataset. In research articles Akram et al. (2019),

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Kamal and Ahuja (2019)	Springer Book	2019	Academic Performance Prediction Using Data Mining Techniques: Identification of Influential Factors Effecting the Academic Performance in Undergrad Professional Course	ID3	–	480 students of BCA.	–	SAS/ Enterprise Miner	Question-naires from students enrolled in Bachelor of Computer Applications (BCA)	–
Dimić et al. (2019)	Springer Book	2019	An Approach to Educational Data Mining Model Accuracy Improvement Using Histogram Discretization and Combining Classifiers into an Ensemble	Naïve Bayes, Hidden Naı́ve Bayes, J48 and Random Forest classifiers	True positive (TP) rate, False positive (FP) rate, Precision	276 records	–	–	Students' activities in e-learning course from Moodle database	–
Zaffar et al. (2018)	Springer Proceeding	2019	Comparing the Performance of FCBF, Chi-Square and Relief-F Filter Feature Selection Algorithms in Educational Data Mining	FCBF, Chi-Square, and ReliefF, Classification - SVM	Accuracy	1. 395 students with 33 attributes.2. 500 students records and 16 features.3. 300 students' record along with 21 features.	–	–	UCI Machine Learning repository and different colleges of India	–
Vila et al. (2018)	Springer Proceeding	2019	Detection of Desertion Patterns in University Students Using Data Mining Techniques:A Case Study	RandomTree, Naïve Bayes	Accuracy, Precision, Recall, Receiver Operating Characteristic, F-measures, Cross-validation K-fold	17,882 student records., personal and academic students data from 37 un-dergraduate careers	2013–2017	Weka	The academic database of the UTN	98%
Almutairi et al. (2019)	Springer Proceeding	2019	Predicting Students' Academic Performance and Main Behavioral Features Using Data Mining Techniques	Random forest , Extreme gradient boosting (XGBoost), Logistic regression, Multilayer perceptron artificial neural network (MLP)	Accuracy, Precision, Recall, F-measures, Cross-validation K-fold	480 students with 16 features	–	–	A LMS called Kalboard 360	75.20%
Kasthuri-arachchi and Liyanage (2018)	Springer Proceeding	2019	Predicting Students' Academic Performance Using Utility Based Educational Data Mining	Naïve Bayes, Random Forest and Decision Tree	Accuracy, Cross-validation K-fold	250 instances with 11 attributes.	–	–	Databases from ques-tionnaires, surveys, interviews, and other databases	98.90%

(*continued on next page*)

Francis and Babu (2019), Santoso (2019), and El Aissaoui et al. (2020), various classification algorithms were compared based on various performance parameters such as Accuracy, Recall, Precision, F-measures, etc. to find the best classification algorithm to be applied on the student dataset. Table 7 gives the analysis of research articles based on clustering algorithm along with dataset used, performance evaluation measures, sampling period, tools/ softwares used and data collection techniques.

3.4. Classification and Association Rule Mining

This section presents that how Classification and Association Rule Mining are used in students' performance prediction. Research (Badr et al., 2016) had made use of classification algorithm followed by association rule algorithm to find the students' performance in programming courses based on the performance in English and Mathematics courses. In research Ayub et al. (2017), J48 classification algorithm

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Ab Ghani et al. (2018)	Springer Proceeding	2019	Student Enrollment Prediction Model in Higher Education Institution: A Data Mining Approach	Logistic regression, Decision tree, Naïve Bayes	Accuracy, Recall, Cross-validation K-fold	the dataset to 1,796 examples for information technology, 3,748 examples for engineering, and 5,328 examples for business management. - 09 attributes	–	–	Student data	71%
Ibrahim et al. (2018)	Springer Proceeding	2019	Mining Unit Feedback to Explore Students' Learning Experiences	Support Vector Machines (SVM), Naive Bayes, Decision Tree, Random Forest	Accuracy, Precision, Recall, F-measures	797 instances of students	2012–2016	KNIME	Hand-written feedback,	73%
Yousafzai et al. (2020)	Springer Journal	2020	Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student	Decision Tree, K-Nearest Neighbour	Accuracy, Cross-validation K-fold, Root mean square error (RMSE)	80,000 historical students' data,	–	–	Student data	96.64%
Adekitan and Salau (2019)	Elsevier Journal	2019	The impact of engineering students' performance in the first three years on their graduation result using educational data mining	Probabilistic Neural Network (PNN) based on the DDA (Dynamic Decay Adjustment), the Random Forest Predictor, the Decision Tree Predictor, the Naïve Bayes Predictor, the Tree Ensemble Predictor, and the Logistic Regression Predictor	Accuracy	1,841 students	2002–2014	KNIME	Student data	89.15%
Yahya (2019)	Elsevier Journal	2019	Swarm intelligence-based approach for educational data classification	Particle Swarm Classification	t-Test	Question data set	–	–	Question data set	–
Lottering et al. (2020)	IEEE Proceeding	2020	A model for the identification of students at risk of **dropout** at a university of technology	Decision Trees (C5.0), Support Vector Machine, Naive Bayes and Nearest Neighbour	Accuracy, Precision, Recall, F-measures	Harvest data from the university database of 4417 full-time students with 19 attributes	2013–2017	R Programming	Harvest data from the university database of full-time students	99%

(*continued on next page*)

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Utari et al. (2020)	IEEE Proceeding	2020	Implementation of Data Mining for Drop-Out Prediction using Random Forest Method	Random Forest	Accuracy, Precision, Recall,Area under Curve (AUC), F-measures, Cross-validation K-fold	2492 student data with 32 attributes	2008–2012.	Java NetBeans IDE 8.2	Student's academic data	92.27%
Mengash (2020)	IEEE Transaction	2020	Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems	Artificial Neural Network, Decision Trees, Support Vector Machines, and Naïve Bayes	Accuracy, Precision, Recall, F-measures, Cross-validation K-fold	2,039 students	2016–2019	–	Students enrolled in a Computer Science and Information College	79.22%
Rahman and Mahmud (2020)	ACM Proceeding	2020	Classification on Educational Performance Evaluation Dataset using Feature Extraction Approach-	K-Nearest Neighbours (KNN), Support Vector Machine (SVM) and Convolutional Neural Network (CNN)	Accuracy, Cross-validation K-fold	16 features and a total of 378 data objects	–	–	Student dataset	56.58%
Mkwazu and Yan (2020)	ACM Proceeding	2020	Grade Prediction Method for University Course Selection Based on Decision Tree	Support Vector Machine (SVM), Decision Trees (DT), and Logistic Regression (LR).		886012 records and 11 attributes.	2014–2019	–	Dataset from the Sokoine University of Agriculture Student Information	–
Islam and Mahmud (2020)	ACM Proceeding	2020	Integration of Learning Analytics into Learner Management System using Machine Learning	Logistic Regression (LR), K-nearest Neighbours (KNN) and Decision Trees (DT).	Accuracy	32593 records relating to the various learner manage-ment activities.	–	–	the 'Open University Learning Analytics' dataset for the experiment.	63%
Injadat et al. (2020a)	Springer Journal	2020	Multi-split optimized bagging ensemble model selection for multi-class educational data mining	SVM-RBF, Logistic Regression, NB, k-NN, Random Forest, NN1, NN2, and NN3	Precision, Recall, F-measure , False Positive Rate	115 first year engi-neering students and event log of the 486 students enrolled.	–	–	Results of a series of tasks performed by University students.	–
Karthikeyan et al. (2020)	Springer Journal	2020	Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation	Hybrid Educational Data Mining model with two classification models -Naïve Bayes and J48	Accuracy, Precision, Recall, F-measures	Student dataset with 14 attributes	2008	Weka	Student results and information under various branches that are collected from the database of the institution.	98.60%
Jung (2018)	Springer Proceeding	2020	An Educational Data Mining with Bayesian Networks for Analyzing Variables Affecting Parental Attachment	Bayesian Network	–	2,564 subjects with 33 variables	2004 and 2008.	R programming	Korean Youth Panel Survey, a longitudinal study of a nationally representative sample of South Korean children and adolescents collected by the National Youth Policy Institute of South Korea	–

(*continued on next page*)

Table 6 (*continued*).

Ref.	Research database	Year	Title	Classification algorithm/s used	Performance evaluation measures	Dataset used	Sampling period	Tool/ Software used	Data collection technique	Accuracy
Agrawal et al. (2020)	Springer Book	2020	Performance Appraisal of an Educational Institute Using Data Mining Techniques	Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), K-Nearest Neighbour (KNN), Support Vector Machines (SVM), and Random Forest (RF)	Confusion Matrix	50 faculties from 379 students	–	R programming	Faculty and student data	–
Ribeiro and Canedo (2020)	Springer Proceeding	2020	Using Data Mining Techniques to Perform School Dropout Prediction: A Case Study	Generalized Linear Model, Gradient Boosting Machine, Support Vector Machines, Random Forest.	Confusion Matrix, Receiver Operating Characteristic (ROC)	data sample with a total of 35,646 students	2006 and 2018	H2O - a Java-based software for data modeling and general computing	Student data	–
Ashraf et al. (2020)	Elsevier Journal	2020	An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering approaches	J48, Random Tree, Naïve Bayes, K-nearest neighbour, and Boosting	True positive rate, True positive rate,Precision, Recall, Receiver Operating Characteristic, F-measures, Correctly Classified Instances, Incorrectly Classified Instances, Cross-validation K-fold, Relative absolute error	115 Students data	–	–	Pedagogical dataset	–
Injadat et al. (2020b)	Elsevier Proceeding	2020	Systematic ensemble model selection approach for educational data mining	K-nearest Neighbour, Naïve Bayes, IBK, J48, Adaboost, Logitboost, PART, Random Forest, Bagging, SMO	Accuracy, Precision, F-measures, Cross-validation K-fold	538 Students data	–	R Programming	LMS data	–

and association rule algorithms were applied on the dataset obtained from Learning Course Management System about two course Change Management and Creative Leadership to improve students' engagement in blended learning mode. It is found in research Sukhija et al. (2018) that association rule algorithm followed by classification algorithm was useful to find the relationship among the association rules. The research Rojanavasu (2019) applied the association rule algorithm to find the data for admission planning. The decision tree classification algorithm was used on students' academic data and placement data of graduated students to predict the job opportunity after graduation (Rojanavasu, 2019). Classification algorithm were used in the relational association rule algorithm for determining students' performance in terms of Pass/Fail. Table 8 presents the analysis of research articles based on classification and association rule mining technique.

3.5. Classification, Clustering and Association rule mining

This section discusses the use of Classification, Clustering and Association rule mining in EDM to analyze the dataset from educational sector. In research Kan et al. (2010), author proposed the Data Mining based Course Management System in which the classification algorithm - Decision Tree, clustering algorithm - K-means and Association rule algorithm – Apriori were considered to predict students' performance. In research Trandafili et al. (2012), the authors Evis Trandafili, et al. applied decision tree classification algorithm on the student dataset to predict students' assessment in form of Pass/ Fail while Expectation Maximization clustering algorithm was used to form the group of students in distinct profiles. Apriori association rule algorithm was considered to discover the relationship among the various academic courses. Research Mayilvaganan and Kalpanadevi (2015) illustrated the use of Naïve Bays classification algorithm, K-means clustering algorithm and FP Growth association rule algorithm using Data Mining tool RapidMiner to assess students' skill on basis of problem solving resources. Table 9 shows the analysis of research articles based on classification, clustering and association rule mining technique with performance evaluation measures, dataset used, tool/software used and data collection technique.

3.6. Classification, Regression, and Clustering

This section presents the use of Classification, Regression, and Clustering in EDM. The research article Jacob et al. (2015), Angra and Ahuja (2017) made the use of these three techniques for predicting students'performance in education sector. In research Jacob et al. (2015), ID3 was applied on the dataset containing students'10th marks, 12th marks, board of study, pass/fail result, backlogs, dropout year, attendance, class test and grade with performance parameters accuracy using Weka tool to predict students' academic performance. It is found that the accuracy was 95%. Multiple regression technique was used on the student dataset containing attendance, presentation score, assignment score, class test score, lab grade scored, average semester score and cumulative grade point average to predict the grade point average. It is noted that 80% of predictions found to be correct as mention in the research article Jacob et al. (2015). K-means clustering algorithm was applied on the dataset to form the groups for placement for helping students to prepare for particular job profile such as Business Analysts, Software Engineer, etc (Jacob et al., 2015). In research Angra and Ahuja (2017), classification algorithm J48, clustering algorithm k-means and linear regression analysis were used on the dataset containing online survey questionnaire data to analyze the students' grade based on previous year grade. Table 10 explains the analysis of research articles based these three techniques.

Table 7
Analysis based on classification and clustering technique.

Ref.	Research database	Year	Title	Performance evaluation measures	Dataset used	Sampling period	Clustering technique used	Methods of cluster	Tool/ Software used	Data collection technique
Bodea et al. (2010)	Springer Book	2010	Student Performance in Online Project Management Courses: A Data Mining Approach	True positive rate, False positive rate, Precision, Recall, Receiver Operating Characteristic, F-measures, Correctly Classified Instances, Incorrectly Classified Instances, Kappa Statistic, Mean Absolute Error, Root mean square error, Relative absolute error, Root relative squared error	182 student data with 6 attributes	–	Simple K-means clustering algorithm	Cluster centroid	Weka	Student data of project manage-ment
Chel-latamilan et al. (2011)	IEEE Pro-ceeding	2011	Effect of Mining educational Data to improve Adaptation of learning in e-Learning System	True positive rate, False positive rate, Precision, Recall, Receiver Operating Characteristic, F-measures, Correctly Classified Instances, Incorrectly Classified Instances, Kappa Statistic, Cross-validation K-fold, Mean Absolute Error, Root mean square error, Relative absolute error , Root relative squared error	66 students of a course "Database Systems"	–	K-means clustering algorithm	Square Error Criterion	KEEL	Web based learning manage-ment system -MOODLE data
Zengin et al. (2011)	Elsevier Proceed-ing	2011	A sample study on applying data mining research techniques in educational science: developing a more meaning of data	t-Test	531 senior university students in seven different depart-ments	2009–2010 fall semester	Microsoft clustering	–	–	Students data
Bresfelean et al. (2012)	Springer Book	2012	Data Mining Tasks in a Student-Oriented DSS	Correctly Classified Instances, Cross-validation K-fold	student data with 23 attributes	–	K-means clustering algorithm	chi-squared statistic	–	Student data
Akram et al. (2019)	IEEE Transac-tion	2019	Predicting Students' Academic Procrastination in Blended Learning Course Using Homework Submission Data	Kappa Statistic, Cross-validation K-fold, Root mean square error (RMSE)	115 students	Spring 2018.	K-means clustering algorithm	standard-ized cluster centroids with different values of k	–	SCHOLAT Course logs data
Francis and Babu (2019)	Springer Journal	2019	Predicting Academic Performance of Students Using a Hybrid Data Mining Approach	Accuracy, Precision, Recall, F-measures	Demo-graphic features, academic features, behavior features and extra features.	–	K-means clustering algorithm	–	–	Student data
M.Á. et al. (2020)	IEEE Transac-tion	2020	Educational Data Mining for Tutoring Suppo0rt in Higher Education: A Web-Based Tool Case Study in engineering Degrees	Accuracy	21,000 graduated and non-graduated students with 22 attributes	2011 to 2017	K-means clustering algorithm	–	Python, HTML5, Java Script and CSS3	

(*continued on next page*)

3.7. Classification, and EnsembleTechnique

The authors - Samuel-Soma M. Ajibade, Nor Bahiah Ahmad, and Siti Mariyam Shamsuddin Ajibade et al. (2018) proposed the students' performance prediction model based on students' behavior feature which was assessed using classification algorithms such as Discriminant Analysis, K-Nearest Neighbour, Naïve Bayesian, Decision Tree, and Pairwise Coupling on the dataset containing academic features, demographic features, parent's involvement in learning process and behavioral features. The ensemble techniques such as AdaBoost and

Table 7 (*continued*).

Ref.	Research database	Year	Title	Performance evaluation measures	Dataset used	Sampling period	Clustering technique used	Methods of cluster	Tool/ Software used	Data collection technique
Santoso (2019)	Springer Book	2019	The Analysis of Student Performance Using Data Mining	Area under Curve (AUC), Cross-validation K-fold	File with 55 attributes and 1665 records of students	2010 to 2014	K-means clustering algorithm	quadratic error of each iteration	Rapid-Miner	The data sources of the academic information system
El Aissaoui et al. (2020)	Springer Book	2020	Mining Learners' Behaviors: An Approach Based on Educational Data Mining Techniques	Correctly Classified Instances, Incorrectly Classified Instances, Kappa Statistic, Cross-validation K-fold, Mean Absolute Error, Root mean square error, Relative absolute error, Root relative squared error (RRSE)	1235 sequences	–	K-Modes clustering	modes using a frequency based method	Weka	E-learning MOODLE platform
Kapgate (2022)	Helion	2021	Unveiling educational patterns at a regional level in Colombia: data from elementary and public high school institutions	F1 Score	180877 records related to students	–	–	Clustering with distance based	Rapid-Miner	Data from three resources such as educational institution, regional and national.

Bagging were used to enhance the accuracy of built model. It is found that the accuracy of model after considering the ensemble techniques was 94.1%. Table 11 shows the details about this article (Ajibade et al., 2018) such as performance evaluation measures, dataset used, tool/software used and data collection technique.

4. Analysis and discussion

This section discusses the analysis research articles based on number of classification articles considered yearwise, classification technique with other Data Mining Technique used in EDM, classifier as per Weka tool, classification techniques, clustering techniques, association rule techniques, selecting the best classification Technique, classification performance metric, software used in EDM, sampling period, dataset size, and data mining tools.

4.1. Yearwise number of research articles employing classification techniquein EDM

This subsection discusses about number of research articles employing classification technique from particular year. Fig. 7 shows the analysis based on number of research articles which used classification technique to analyze students dataset, from year 2010 to 2020. From Fig. 7, it is observed that maximum number of research articles that is 25 are from year 2019 while only one research articles is considered from year 2010. Sixteen research articles are from year 2015 and 2018 which employed classification technique on students dataset. In year 2010 and 2013, ten research articles used classification technique to predict students' performance. Fourteen research articles Yousafzai et al. (2020), Lottering et al. (2020), Utari et al. (2020), Mengash (2020), Rahman and Mahmud (2020), Mkwazu and Yan (2020), Islam and Mahmud (2020), Injadat et al. (2020a), Karthikeyan et al. (2020), Jung (2018), Agrawal et al. (2020), Ribeiro and Canedo (2020), Ashraf et al. (2020), Injadat et al. (2020b) are from year 2020.

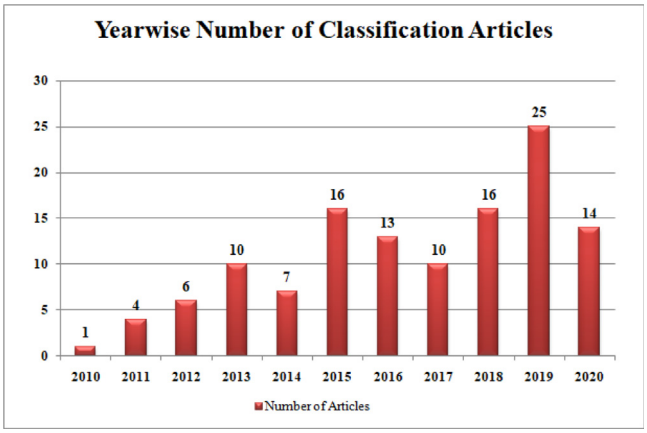


Fig. 7. Yearwise analysis of number of research articles employing classification technique.

4.2. Analysis on basis of classification with other data mining technique used in EDM

Analysis of research articles in terms of classification technique with other Data Mining techniques such as clustering, association rule algorithms, regression and ensemble technique is presented in this subsection. Nine research articles Bodea et al. (2010), Chellatamilan et al. (2011), Zengin et al. (2011), Bresfelean et al. (2012), Akram et al. (2019), Francis and Babu (2019), M.Á. et al. (2020), Santoso (2019), and El Aissaoui et al. (2020) employed both classification and clustering techniques to analyze the students dataset while the algorithms related to the classification and association rule mining are considered in the research articles Badr et al. (2016), Ayub et al. (2017), Sukhija et al. (2018), Rojanavasu (2019), and Crivei et al. (2019). Three Data Mining techniques such as classification, clustering and association rule algorithm are used in the research articles Kan et al. (2010), Trandafili et al. (2012), and Mayilvaganan and Kalpanadevi (2015) while classification, regression, and clustering algorithms are considered in research

Table 8
Analysis based on classification and association rule mining technique.

Ref.	Research database	Year	Title	Performance evaluation measures	Dataset used	Sampling Period	Tool/ Software used	Data collection technique
Badr et al. (2016)	Elsevier Proceeding	2016	Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department	Accuracy	203 students data	–	Java	Students data
Ayub et al. (2017)	IEEE Proceeding	2017	Modeling Online Assessment in Management Subjects through Educational Data Mining	Accuracy, Cross-validation K-fold	314 student data	2014–2015, 2015–2016, and 2016–2017	Weka	Student data
Sukhija et al. (2018)	Springer Book	2018	EDARC: Collaborative Frequent Pattern and Analytical Mining Tool for Exploration of Educational Information	–	2 million student records regarding examination, personal, and infrastructure details	2011–2016	–	Student dataset of matriculation and senior secondary courses
Rojanavasu (2019)	IEEE Proceeding	2019	Educational Data Analytics using Association Rule Mining and Classification	Accuracy, Cross-validation K-fold	1. 10,342 records. 2. 106 records. student's course grade	1. 2016 and 2017 2. 2011–2014	Rapid-Miner	Student's course grade of school of information and communication technology
Crivei et al. (2019)	Springer Proceeding	2019	A Study on Applying Relational Association Rule Mining Based Classification for Predicting the Academic Performance of Students	True positive rate, False positive rate, False Negatives Rate, True Negatives rate, Area under Curve	1. 384 instances 2. 863 instances. 3. 1154 instances	2016–2017, 2014–2018, 2012–2018	–	Three real academic data sets

Table 9
Analysis based on classification, clustering and association rule mining technique.

Ref.	Research database	Year	Title	Performance evaluation measures	Dataset used	Clustering technique used	Association rule technique used	Tool/ Software used	Data collection technique
Kan et al. (2010)	IEEE Proceeding	2010	DMCMS: A Data Mining Based Course Management System	Association Rule - Confidence, Classification - Probability and Clustering - number of cluster	117 students	K-means	Apriori association rule	–	Student data during their learning of the course Decision Analysis
Trandafili et al. (2012)	ACM Proceeding	2012	Discovery and evaluation of student's profiles with machine learning	–	35 000 rows	Expectation Maximization	Apriori association rule	Weka	Student's database from the information system
Mayilvaganan and Kalpanadevi (2015)	Elsevier Proceeding	2015	Cognitive Skill Analysis for Students through Problem Solving Based on Data Mining Techniques	True positive rate, False positive rate, Precision, Recall, Receiver Operating Characteristic, F-measures, Correctly Classified Instances, Incorrectly Classified Instances	200 with 10 attributes Student data	K-means	FP Growth association rule	–	Student data through test based on cognitive model

Table 10
Analysis based on classification, regression and clustering technique.

Ref.	Research database	Year	Title	Performance evaluation measures	Dataset used	Clustering technique used	Regression technique used	Tool/ Software used	Data collection technique	Complexity
Jacob et al. (2015)	IEEE Proceeding	2015	Educational Data Mining Techniques and their Applications	Accuracy, Cross-validation K-fold	Student dataset	K-means	Multiple regression analysis	Weka	Student dataset	80% prediction correct
Angra and Ahuja (2017)	IEEE Proceeding	2017	Implementation of Data Mining Algorithms on Student's data using RapidMiner	–	Student dataset	K-means	Linear egression analysis	RapidMiner	Online survey based on questionnaire	–

Table 11
Analysis based on classification, and ensemble technique.

Ref.	Research database	Year	Title	Performance evaluation measures	Dataset used	Ensemble technique used	Tool/ Software used	Data collection technique
Ajibade et al. (2018)	Sringer Proceed-ing	2020	A Data Mining Approach to Predict Academic Performance of Students Using Ensemble Techniques	Accuracy, Precision, Recall, F-measures, Cross-validation K-fold	500 students with 16 features.	AdaBoost and Bagging	MATLAB	Gathering of data from Learning management system - Kalboard 360 e-Learning system

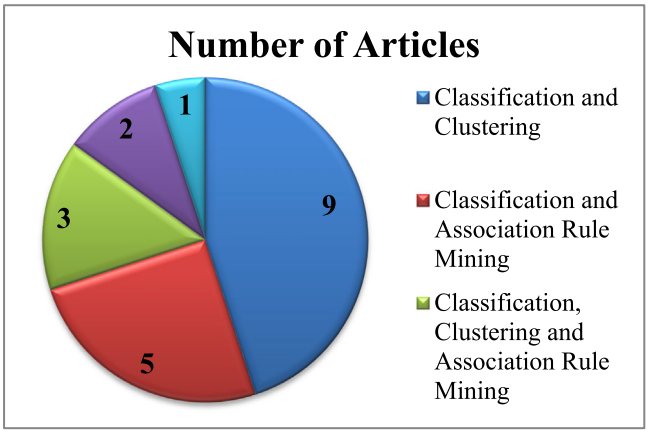


Fig. 8. Analysis on basis of classification with other data mining technique.

articles Jacob et al. (2015), and Angra and Ahuja (2017) to predict the performance of students (see Fig. 8).

4.3. Analysis on basis of classifier as per Weka Tool

This subsection presents the analysis of research articles based on the classifier considered in Weka tool that is seven classifiers – Bays, Function, Lazy, Meta, Rules, Tree and PMML classifier. Table 12 illustrates the analysis in terms of classifiers. It is noted from Table 11 that the mostly considered classifier is Tree which is used by 87 number of research articles (61.27% of total 142 research articles). The classifier Bays which contain two classification algorithms Naïve Bays and Bays Net, is referred by 68 research articles (47.88% of total 142 research articles). Least used classifier is Meta which is referred by the research articles Cocea and Weibelzahl (2010), Dejaeger et al. (2012), Barbosa Manhães et al. (2015), Pise and Kulkarni (2017), Lagus et al. (2018), Spatiotis et al. (2018), Abyaa et al. (2018), and Ajibade et al. (2018). The classifiers Function, Lazy, Rules, and PMML (Predictive Model Markup Language) classifier related algorithms are considered in 47, 26, 20, and 46 number of research articles respectively.

4.4. Analysis on basis of classification techniques

Analysis of research articles in terms of classification algorithm is described in this subsection and in Table 13. From Table 13, it is observed that mostly used classification algorithm is Naïve Bays (44.37% of total 142 research articles) while least used algorithm are Multiple Regression, Kstar, QuickRules Fuzzy-Rough ruleinduction, Fuzzy Nearest Neighbour, Fuzzy Rough Nearest Neighbour, Vaguely Quantified Nearest Neighbour (VQNN), Attribute Selected Classifier, Classification Vis Regression, Conjuctive Rules, M5P, Oblique Classifier, XGBoost, Gaussian Processes Random Tree (GPRT), Generalized Linear Model, Singular Value Decomposition (SVD), Interpretable Classification Rule Mining, MaxMargin Multi-Label (M3L) classifier, LIBSVM, Cross-out classifier, Classifier based on polynomial regression and stochastic gradient descent, and Improved ID3. Support vector machine algorithm is referred in 40 number of research articles (28.16% of total 142 research articles). The classification algorithms - Bays Net, Logistic, Multilayer Perceptron (MLP), K-Nearest Neighbour (KNN), Jrip, J48, Random Forest, ID3, C4.5, Neural Network, and Decision Tree are considered in 12, 21, 16, 21, 11, 24, 38, 10, 22, 15 and 26 number of research articles respectively.

4.5. Analysis on basis of clustering techniques

Analysis on basis of clustering techniques is presented in this subsection. Clustering algorithms such as K-means, Expectation Maximization, and K-modes are considered in the research articles. It is found from Table 14 that K-means clustering algorithm is used by the research articles Kan et al. (2010), Bodea et al. (2010), Chellatamilan et al. (2011), Bresfelean et al. (2012), Jacob et al. (2015), Mayilvaganan and Kalpanadevi (2015), Angra and Ahuja (2017), Akram et al. (2019), Francis and Babu (2019), and M.Á. et al. (2020) while Expectation Maximization, and K-modes algorithms are considered by the research articles Trandafili et al. (2012), El Aissaoui et al. (2020) respectively.

4.6. Analysis on basis of association rule techniques

This subsection elucidates the analysis of research articles employing Association Rule Techniques in EDM. Table 15 represents

Table 12
Analysis based on classifier considered in WEKA tool.

Sr. No.	Classifier	Number of articles	Reference number
1	Bays	68	Cocea and Weibelzahl (2010) , El-Halees (2011), Chau and Phung (2013), Pratiwi (2013), Göker et al. (2013), Mashiloane and Mchunu (2013), Palazuelos et al. (2013), Dangi and Srivastava (2014), Anh et al. (2014), Chen et al. (2014), Ragab et al. (2014), Manhães et al. (2014), Pruthi and Bhatia (2015), Guo et al. (2015), Guarín et al. (2015), Ahadi et al. (2015), Bakaric et al. (2015), Barbosa Manhães et al. (2015), Jishan et al. (2015), Salinas and Stephens (2015), Kaur et al. (2015), Mayilvaganan and Kalpanadevi (2015), Amornsinlaphachai (2016), Devasia et al. (2016), Lehr et al. (2016), Chaudhury et al. (2016), Sanchez-Santillan et al. (2016), Ahmed et al. (2016), Athani et al. (2017), Castro-Wunsch et al. (2017), Daud et al. (2017), Pise and Kulkarni (2017), Costa et al. (2017), Kiu (2018), Srivastava et al. (2018), Chitra and Agrawal (2019), Niu et al. (2018), Rustia et al. (2018), Abyaa et al. (2018), Chanlekha and Niramitranon (2018), Kularbphettong (2017), Pérez et al. (2018), Maitra et al. (2018), Tasnim et al. (2019), Al Breiki et al. (2019), Lagman et al. (2019), Umer et al. (2019), Amazona and Hernandez (2019), Al Fanah and Ansari (2019), Francis and Babu (2019), Adekitan and Salau (2020), Tarmizi et al. (2019), Rawat and Malhan (2019), Dimić et al. (2019), Vila et al. (2018), Ab Ghani et al. (2018), Ibrahim et al. (2018), Adekitan and Salau (2019), Lottering et al. (2020), Mengash (2020), Santoso (2019), Injadat et al. (2020a), Karthikeyan et al. (2020), Ajibade et al. (2018), Jung (2018), El Aissaoui et al. (2020), Ashraf et al. (2020), Injadat et al. (2020b)
2	Function	47	Cocea and Weibelzahl (2010), Chuan et al. (2011), Wang and Liao (2011), Nasiri et al. (2012), Dejaeger et al. (2012), Sen and Ucar (2012), Chau and Phung (2013), Pratiwi (2013), Göker et al. (2013), Ragab et al. (2014), Manhães et al. (2014), Jacob et al. (2015), Guo et al. (2015), Sorour et al. (2015), Barbosa Manhães et al. (2015), Kaur et al. (2015), Amornsinlaphachai (2016), Lehr et al. (2016), Agaoglu (2016), Hassan and Al-Razgan (2016), Ahmed et al. (2016), Leppänen et al. (2017), Pise and Kulkarni (2017), Jung (2016), Kiu (2018), Chitra and Agrawal (2019), Rustia et al. (2018), Abyaa et al. (2018), Chanlekha and Niramitranon (2018), Pérez et al. (2018), Burgos et al. (2018), Tasnim et al. (2019), Al Breiki et al. (2019), Alshehri (2019), Umer et al. (2019), Altaf et al. (2019), Al Fanah and Ansari (2019), Rawat and Malhan (2019), Almutairi et al. (2019), Ab Ghani et al. (2018), Adekitan and Salau (2019), Mengash (2020), Mkwazu and Yan (2020), Islam and Mahmud (2020), Injadat et al. (2020a), Agrawal et al. (2020), Injadat et al. (2020b)
3	Lazy	26	Cocea and Weibelzahl (2010), Chuan et al. (2011), El-Halees (2011), Chau and Phung (2013), Pratiwi (2013), Anh et al. (2014), Ragab et al. (2014), Auddy and Mukhopadhyay (2015), Amornsinlaphachai (2016), Lehr et al. (2016), Pise and Kulkarni (2017), Srivastava et al. (2018), Spatiotis et al. (2018), Niu et al. (2018), Abyaa et al. (2018), Al Breiki et al. (2019), Rawat and Malhan (2019), Yousafzai et al. (2020), Lottering et al. (2020), Rahman and Mahmud (2020), Islam and Mahmud (2020), Injadat et al. (2020a), Ajibade et al. (2018), Agrawal et al. (2020), Ashraf et al. (2020), Injadat et al. (2020b)
4	Meta	8	Cocea and Weibelzahl (2010), Dejaeger et al. (2012), Barbosa Manhães et al. (2015), Pise and Kulkarni (2017), Lagus et al. (2018), Spatiotis et al. (2018), Abyaa et al. (2018), Ajibade et al. (2018)
5	Rules	20	Bodea et al. (2010), Pratiwi (2013), Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Mashiloane and Mchunu (2013), Palazuelos et al. (2013), Ragab et al. (2014), Manhães et al. (2014), Ahadi et al. (2015), Sisovic et al. (2015), Bakaric et al. (2015), Barbosa Manhães et al. (2015), Auddy and Mukhopadhyay (2015), Amornsinlaphachai (2016), Sanchez-Santillan et al. (2016), Meedech et al. (2016), Castro-Wunsch et al. (2017), Pise and Kulkarni (2017), Al Breiki et al. (2019), Chango et al. (2019)
6	Tree	87	Bodea et al. (2010), Chellatamilan et al. (2011), Cocea and Weibelzahl (2010), Bunkar et al. (2012), Nasiri et al. (2012), Trandafili et al. (2012), Bresfelean et al. (2012), Dejaeger et al. (2012), Şen et al. (2012), Sen and Ucar (2012), Chau and Phung (2013), Pratiwi (2013), Göker et al. (2013), Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Mashiloane and Mchunu (2013), Palazuelos et al. (2013), Pathan et al. (2014), Anh et al. (2014), Ragab et al. (2014), Natek and Zwilling (2014), Jacob et al. (2015), Parmar et al. (2015), Ahadi et al. (2015), Sisovic et al. (2015), Bakaric et al. (2015), Barbosa Manhães et al. (2015), Jishan et al. (2015), Auddy and Mukhopadhyay (2015), Shukor et al. (2015), Kaur et al. (2015), Amornsinlaphachai (2015), Amornsinlaphachai (2016), Lehr et al. (2016), Agaoglu (2016), Chaudhury et al. (2016), Sanchez-Santillan et al. (2016), Ramanathan et al. (2016), Meedech et al. (2016), Hamsa et al. (2016), Ahmed et al. (2016), Angra and Ahuja (2017), Ayub et al. (2017), Castro-Wunsch et al. (2017), Figueira (2017), Daud et al. (2017), Pise and Kulkarni (2017), Kiu (2018), Patil et al. (2018), Kaunang and Rotikan (2018), Chitra and Agrawal (2019), Lagus et al. (2018), Spatiotis et al. (2018), Rustia et al. (2018), Abyaa et al. (2018), Chanlekha and Niramitranon (2018), Martínez-Abad et al. (2018), Kularbphettong (2017), Pérez et al. (2018), Miguéis et al. (2018), Rojanavasu (2019), Ketui et al. (2019), Al Breiki et al. (2019), Umer et al. (2019), Chango et al. (2019), Al Fanah and Ansari (2019), Adekitan and Salau (2020), Tarmizi et al. (2019), Martins et al. (2019), Rawat and Malhan (2019), Kamal and Ahuja (2019), Dimić et al. (2019), Vila et al. (2018), Almutairi et al. (2019), Kasthuriarachchi and Liyanage (2018), Ibrahim et al. (2018), Adekitan and Salau (2019), Lottering et al. (2020), Utari et al. (2020), Santoso (2019), Injadat et al. (2020a), Karthikeyan et al. (2020), El Aissaoui et al. (2020), Agrawal et al. (2020), Ribeiro and Canedo (2020), Ashraf et al. (2020), Injadat et al. (2020b)
7	PMML Classifier	46	Cocea and Weibelzahl (2010), El-Halees (2011), Dejaeger et al. (2012), Chau and Phung (2013), Blagojević and Micić (2013), Anh et al. (2014), Chen et al. (2014), Manhães et al. (2014), Guo et al. (2015), Sorour et al. (2015), Bakaric et al. (2015), Barbosa Manhães et al. (2015), Jishan et al. (2015), Agaoglu (2016), Chaudhury et al. (2016), Castro-Wunsch et al. (2017), Buenaño-Fernández et al. (2017), Kitanaka et al. (2017), Daud et al. (2017), Leppänen et al. (2017), Costa et al. (2017), Srivastava et al. (2018), Chitra and Agrawal (2019), Lagus et al. (2018), Spatiotis et al. (2018), Niu et al. (2018), Rustia et al. (2018), Abyaa et al. (2018), Chanlekha and Niramitranon (2018), Tasnim et al. (2019), Amazona and Hernandez (2019), Francis and Babu (2019), Adekitan and Salau (2020), Tarmizi et al. (2019), Zaffar et al. (2018), Ibrahim et al. (2018), Adekitan and Salau (2019), Lottering et al. (2020), M.Á. et al. (2020), Mengash (2020), Rahman and Mahmud (2020), Mkwazu and Yan (2020), Injadat et al. (2020a), Agrawal et al. (2020), Ribeiro and Canedo (2020), Injadat et al. (2020b)
8	Other	39	Kan et al. (2010), Guruler et al. (2010), Chuan et al. (2011), Zengin et al. (2011), Sakurai et al. (2012), Sen and Ucar (2012), Hoe et al. (2013), Márquez-Vera et al. (2013a), Blagojević and Micić (2013), Chen et al. (2014), Ragab et al. (2014), Gera and Goel (2015), Pruthi and Bhatia (2015), Guarín et al. (2015), Lehr et al. (2016), Agaoglu (2016), Ramanathan et al. (2016), Stahovich and Lin (2016), Kassak et al. (2016), Jung (2016), Costa et al. (2017), Patil et al. (2018), Kaunang and Rotikan (2018), Chanlekha and Niramitranon (2018), Pérez et al. (2018), Ketui et al. (2019), Al Breiki et al. (2019), Amazona and Hernandez (2019), Francis and Babu (2019), Kasthuriarachchi and Liyanage (2018), Ab Ghani et al. (2018), Ibrahim et al. (2018), Yousafzai et al. (2020), Adekitan and Salau (2019), Mengash (2020), Mkwazu and Yan (2020), Islam and Mahmud (2020), Ajibade et al. (2018), Ribeiro and Canedo (2020)

the analysis on basis of association rule techniques. It is noted from Table 15 that Apriori association rule algorithm is employed in the research article Kan et al. (2010), Trandafili et al. (2012), and Rojanavasu (2019) while FP-growth and Relational Association rules are used by the research articles Mayilvaganan and Kalpanadevi (2015), Jung (2018) respectively.

4.7. Analysis of classification performance metrics

Performance parameters are used to decide the performance of different classification techniques. This subsection deals with the analysis of research articles based on classification performance metrics. Table 16 shows the analysis of research articles on basis of 24 performance parameters such as Accuracy, True positive rate, False positive rate, False Negatives, True Negatives, Precision, Recall,

Table 13
Analysis based on classification techniques.

Classification algorithms	Number of articles	Reference number
Naïve Bayes	63	El-Halees (2011), Chau and Phung (2013), Pratiwi (2013), Göker et al. (2013), Mashiloane and Mchunu (2013), Palazuelos et al. (2013), Dangi and Srivastava (2014), Anh et al. (2014), Chen et al. (2014), Ragab et al. (2014), Manhães et al. (2014), Pruthi and Bhatia (2015), Guo et al. (2015), Guarín et al. (2015), Ahadi et al. (2015), Bakaric et al. (2015), Barbosa Manhães et al. (2015), Jishan et al. (2015), Salinas and Stephens (2015), Kaur et al. (2015), Mayilvaganan and Kalpanadevi (2015), Amornsinlaphachai (2016), Devasia et al. (2016), Lehr et al. (2016), Chaudhury et al. (2016), Ahmed et al. (2016), Athani et al. (2017), Castro-Wunsch et al. (2017), Daud et al. (2017), Pise and Kulkarni (2017), Costa et al. (2017), Kiu (2018), Srivastava et al. (2018), Chitra and Agrawal (2019), Niu et al. (2018), Rustia et al. (2018), Abyaa et al. (2018), Chanlekha and Niramitranon (2018), Pérez et al. (2018), Maitra et al. (2018), Tasnim et al. (2019), Al Breiki et al. (2019), Lagman et al. (2019), Umer et al. (2019), Amazona and Hernandez (2019), Francis and Babu (2019), Adekitan and Salau (2020), Tarmizi et al. (2019), Rawat and Malhan (2019), Dimić et al. (2019), Vila et al. (2018), Ab Ghani et al. (2018), Ibrahim et al. (2018), Adekitan and Salau (2019), Lottering et al. (2020), Mengash (2020), Santoso (2019), Injadat et al. (2020a), Karthikeyan et al. (2020), Ajibade et al. (2018), El Aissaoui et al. (2020), Ashraf et al. (2020), Injadat et al. (2020b)
Bays Net	12	Cocea and Weibelzahl (2010), Göker et al. (2013), Palazuelos et al. (2013), Ahadi et al. (2015), Barbosa Manhães et al. (2015), Amornsinlaphachai (2016), Sanchez-Santillan et al. (2016), Daud et al. (2017), Kularbphettong (2017), Al Breiki et al. (2019), Al Fanah and Ansari (2019), Jung (2018)
Logistic	21	Cocea and Weibelzahl (2010), Chuan et al. (2011), Chau and Phung (2013), Barbosa Manhães et al. (2015), Lehr et al. (2016), Jung (2016), Rustia et al. (2018), Abyaa et al. (2018), Pérez et al. (2018), Burgos et al. (2018), Tasnim et al. (2019), Alshehri (2019), Umer et al. (2019), Al Fanah and Ansari (2019), Almutairi et al. (2019), Ab Ghani et al. (2018), Adekitan and Salau (2019), Mkwazu and Yan (2020), Islam and Mahmud (2020), Injadat et al. (2020a), Injadat et al. (2020b)
Simple Logistic	2	Cocea and Weibelzahl (2010), Al Breiki et al. (2019)
Linear Discriminating analysis	2	Umer et al. (2019), Agrawal et al. (2020)
Multilayer Perceptron (MLP)	16	Dejaeger et al. (2012), Ragab et al. (2014), Manhães et al. (2014), Guo et al. (2015), Barbosa Manhães et al. (2015), Kaur et al. (2015), Lehr et al. (2016), Ahmed et al. (2016), Jung (2016), Kiu (2018), Chitra and Agrawal (2019), Al Breiki et al. (2019), Altaf et al. (2019), Rawat and Malhan (2019), Almutairi et al. (2019), Injadat et al. (2020b)
Artificial Neural Network	7	Wang and Liao (2011), Sen and Ucar (2012), Sorour et al. (2015), Amornsinlaphachai (2016), Agaoglu (2016), Chanlekha and Niramitranon (2018), Mengash (2020)
Sequential minimal optimization (SMO)	5	Pratiwi (2013), Kaur et al. (2015), Ahmed et al. (2016), Pise and Kulkarni (2017), Al Breiki et al. (2019)
Linear Regression	3	Nasiri et al. (2012), Hassan and Al-Razgan (2016), Al Breiki et al. (2019)
Multiple Regression	1	Jacob et al. (2015)
RBF Network	3	Göker et al. (2013), Manhães et al. (2014), Leppänen et al. (2017)
IBk (Instance based Learner)	4	Cocea and Weibelzahl (2010), Ragab et al. (2014), Pise and Kulkarni (2017), Spatiotis et al. (2018)
K-Nearest Neighbour (KNN)	21	Chuan et al. (2011), El-Halees (2011), Chau and Phung (2013), Anh et al. (2014), Amornsinlaphachai (2016), Lehr et al. (2016), Pise and Kulkarni (2017), Srivastava et al. (2018), Niu et al. (2018), Abyaa et al. (2018), Al Breiki et al. (2019), Rawat and Malhan (2019), Yousafzai et al. (2020), Lottering et al. (2020), Rahman and Mahmud (2020), Islam and Mahmud (2020), Injadat et al. (2020a), Ajibade et al. (2018), Agrawal et al. (2020), Ashraf et al. (2020), Injadat et al. (2020b)
Kstar	1	Pratiwi (2013)
QuickRules Fuzzy-Rough ruleinduction	1	Auddy and Mukhopadhyay (2015)
Fuzzy Nearest Neighbour	1	Auddy and Mukhopadhyay (2015)
Fuzzy Rough Nearest Neighbour	1	Auddy and Mukhopadhyay (2015)
Vaguely Quantified Nearest Neighbour (VQNN)	1	Auddy and Mukhopadhyay (2015)
AdaBoost	5	Barbosa Manhães et al. (2015), Pise and Kulkarni (2017), Lagus et al. (2018), Spatiotis et al. (2018), Ajibade et al. (2018)
Attribute Selected Classifier	1	Cocea and Weibelzahl (2010)
Bagging	5	Cocea and Weibelzahl (2010), Pise and Kulkarni (2017), Spatiotis et al. (2018), Abyaa et al. (2018), Ajibade et al. (2018)
Classification Vis Regression	1	Cocea and Weibelzahl (2010)
Logit Boot	2	Dejaeger et al. (2012), Pise and Kulkarni (2017)
Decision Table	6	Mashiloane and Mchunu (2013), Manhães et al. (2014), Ahadi et al. (2015), Barbosa Manhães et al. (2015), Castro-Wunsch et al. (2017), Al Breiki et al. (2019)
Conjuctive Rules	1	Ahadi et al. (2015)
Jrip	11	Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Palazuelos et al. (2013), Sisovic et al. (2015), Barbosa Manhães et al. (2015), Auddy and Mukhopadhyay (2015), Amornsinlaphachai (2016), Sanchez-Santillan et al. (2016), Meedech et al. (2016), Al Breiki et al. (2019), Chango et al. (2019)
OneR	5	Pratiwi (2013), Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Barbosa Manhães et al. (2015), Meedech et al. (2016)
PART	8	Bodea et al. (2010), Ragab et al. (2014), Ahadi et al. (2015), Sisovic et al. (2015), Bakaric et al. (2015), Castro-Wunsch et al. (2017), Pise and Kulkarni (2017), Chango et al. (2019)
Prism	2	Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a)
RIDOR	3	Márquez-Vera et al. (2013a), Palazuelos et al. (2013), Meedech et al. (2016)
Nnge	3	Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Chango et al. (2019)
Decision Stump	2	Ahadi et al. (2015), Castro-Wunsch et al. (2017)

(*continued on next page*)

Confusion Matrix, Area under Curve, Receiver Operating Characteristic, F-measures, Correctly Classified Instances, Incorrectly Classified Instances, Kappa Statistic, Cross-validation K-fold, Mean Absolute Error, Root mean square error, Relative absolute error, Root relative squared error, Notch difference, Lift Value, Standard Deviation, t-Test, and Cosine Cofficient. From Table 15, it is found that the performance parameter — accuracy is used by 84 research articles such as Cocea and Weibelzahl (2010), Nasiri et al. (2012), Dejaeger et al. (2012), Şen et al. (2012), Sen and Ucar (2012), Hoe et al. (2013), Chau and Phung (2013), Pratiwi (2013), Göker et al. (2013), Márquez-Vera

Table 13 (*continued*).

Classification algorithms	Number of articles	Reference number
J48	24	Bodea et al. (2010), Chellatamilan et al. (2011), Cocea and Weibelzahl (2010), Trandafili et al. (2012), Pratiwi (2013), Göker et al. (2013), Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Mashiloane and Mchunu (2013), Palazuelos et al. (2013), Natek and Zwilling (2014), Parmar et al. (2015), Ahadi et al. (2015), Sisovic et al. (2015), Barbosa Manhães et al. (2015), Auddy and Mukhopadhyay (2015), Kaur et al. (2015), Sanchez-Santillan et al. (2016), Meedech et al. (2016), Ahmed et al. (2016), Angra and Ahuja (2017), Ayub et al. (2017), Castro-Wunsch et al. (2017), Pise and Kulkarni (2017), Kiu (2018), Chitra and Agrawal (2019), Spatiotis et al. (2018), Abyaa et al. (2018), Martínez-Abad et al. (2018), Kularbphettong (2017), Chango et al. (2019), Tarmizi et al. (2019), Rawat and Malhan (2019), Dimić et al. (2019), Karthikeyan et al. (2020), Ashraf et al. (2020)
M5P	1	Natek and Zwilling (2014)
ADTree	4	Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Ahadi et al. (2015), Meedech et al. (2016)
Random Forest	38	Chau and Phung (2013), Márquez-Vera et al. (2013b), Palazuelos et al. (2013), Anh et al. (2014), Ragab et al. (2014), Parmar et al. (2015), Ahadi et al. (2015), Barbosa Manhães et al. (2015), Lehr et al. (2016), Meedech et al. (2016), Castro-Wunsch et al. (2017), Figueira (2017), Pise and Kulkarni (2017), Kiu (2018), Kaunang and Rotikan (2018), Chitra and Agrawal (2019), Lagus et al. (2018), Spatiotis et al. (2018), Abyaa et al. (2018), Chanlekha and Niramitranon (2018), Pérez et al. (2018), Miguéis et al. (2018), Al Breiki et al. (2019), Umer et al. (2019), Al Fanah and Ansari (2019), Adekitan and Salau (2020), Tarmizi et al. (2019), Martins et al. (2019), Dimić et al. (2019), Almutairi et al. (2019), Kasthuriarachchi and Liyanage (2018), Ibrahim et al. (2018), Adekitan and Salau (2019), Utari et al. (2020), Injadat et al. (2020a), Agrawal et al. (2020), Ribeiro and Canedo (2020), Injadat et al. (2020b)
Random Tree	6	Márquez-Vera et al. (2013a), Ramanathan et al. (2016), Ketui et al. (2019), Chango et al. (2019), Vila et al. (2018), Ashraf et al. (2020)
REPTree	7	Bresfelean et al. (2012), Márquez-Vera et al. (2013a), Natek and Zwilling (2014), Kaur et al. (2015), Meedech et al. (2016), Spatiotis et al. (2018), Chango et al. (2019)
CART (Classification and Regression Trees)	7	Bunkar et al. (2012), Dejaeger et al. (2012), Agaoglu (2016), Daud et al. (2017), Spatiotis et al. (2018), El Aissaoui et al. (2020), Agrawal et al. (2020)
SimpleCart	4	Pratiwi (2013), Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Meedech et al. (2016)
ID3	10	Bunkar et al. (2012), Pathan et al. (2014), Jacob et al. (2015), Amornsinlaphachai (2015), Amornsinlaphachai (2016), Patil et al. (2018), Rojanavasu (2019), Ketui et al. (2019), Kamal and Ahuja (2019), El Aissaoui et al. (2020)
C4.5	22	Cocea and Weibelzahl (2010), Bunkar et al. (2012), Bresfelean et al. (2012), Sen and Ucar (2012), Chau and Phung (2013), Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Pathan et al. (2014), Anh et al. (2014), Ragab et al. (2014), Bakaric et al. (2015), Jishan et al. (2015), Shukor et al. (2015), Amornsinlaphachai (2016), Agaoglu (2016), Chaudhury et al. (2016), Hamsa et al. (2016), Daud et al. (2017), Patil et al. (2018), Rustia et al. (2018), Santoso (2019), El Aissaoui et al. (2020)
C5.0	3	Nasiri et al. (2012), Şen et al. (2012), Lottering et al. (2020)
Oblique Classifier	1	Dejaeger et al. (2012)
XGBoost	1	Almutairi et al. (2019)
Support Vector Machine (SVM)	40	Cocea and Weibelzahl (2010), El-Halees (2011), Dejaeger et al. (2012), Chau and Phung (2013), Anh et al. (2014), Chen et al. (2014), Manhães et al. (2014), Guo et al. (2015), Sorour et al. (2015), Bakaric et al. (2015), Barbosa Manhães et al. (2015), Agaoglu (2016), Chaudhury et al. (2016), Buenaño-Fernández et al. (2017), Kitanaka et al. (2017), Daud et al. (2017), Leppänen et al. (2017), Costa et al. (2017), Srivastava et al. (2018), Chitra and Agrawal (2019), Lagus et al. (2018), Spatiotis et al. (2018), Niu et al. (2018), Rustia et al. (2018), Abyaa et al. (2018), Chanlekha and Niramitranon (2018), Tasnim et al. (2019), Francis and Babu (2019), Tarmizi et al. (2019), Zaffar et al. (2018), Ibrahim et al. (2018), Lottering et al. (2020), M.Á. et al. (2020), Mengash (2020), Rahman and Mahmud (2020), Mkwazu and Yan (2020), Injadat et al. (2020a), Agrawal et al. (2020), Ribeiro and Canedo (2020), Injadat et al. (2020b)
Neural Network	15	Chau and Phung (2013), Blagojević and Micić (2013), Anh et al. (2014), Jishan et al. (2015), Castro-Wunsch et al. (2017), Costa et al. (2017), Spatiotis et al. (2018), Niu et al. (2018), Rustia et al. (2018), Amazona and Hernandez (2019), Francis and Babu (2019), Adekitan and Salau (2020), Adekitan and Salau (2019), Rahman and Mahmud (2020), Injadat et al. (2020a)
Decision Tree (not mentioned Decision Tree algorithm)	26	Kan et al. (2010), Guruler et al. (2010), Zengin et al. (2011), Sakurai et al. (2012), Sen and Ucar (2012), Blagojević and Micić (2013), Gera and Goel (2015), Pruthi and Bhatia (2015), Guarín et al. (2015), Lehr et al. (2016), Ramanathan et al. (2016), Costa et al. (2017), Kaunang and Rotikan (2018), Chanlekha and Niramitranon (2018), Pérez et al. (2018), Amazona and Hernandez (2019), Francis and Babu (2019), Kasthuriarachchi and Liyanage (2018), Ab Ghani et al. (2018), Ibrahim et al. (2018), Yousafzai et al. (2020), Adekitan and Salau (2019), Mengash (2020), Mkwazu and Yan (2020), Islam and Mahmud (2020), Ajibade et al. (2018)
Gradient boosted tree	2	Ketui et al. (2019), Ribeiro and Canedo (2020)
Gaussian Processes Random Tree (GPRT)	1	Al Breiki et al. (2019)
Classification using CHAID (Chi-Squared Automatic Interaction Detection)	2	Hoe et al. (2013), Jung (2016)
Discriminant analysis	2	Agaoglu (2016), Ajibade et al. (2018)
GLM: Generalized Linear Model	1	Ribeiro and Canedo (2020)
Singular Value Decomposition (SVD)	1	Chuan et al. (2011)
Interpretable Classification Rule Mining	1	Márquez-Vera et al. (2013a)
MaxMargin Multi-Label (M3L) classifier	1	Chen et al. (2014)
LIBSVM	1	Ragab et al. (2014)
Cross-out classifier	1	Stahovich and Lin (2016)
Classifier based on polynomial regression and stochastic gradient descent	1	Kassak et al. (2016)
Improved ID3	1	Patil et al. (2018)

Table 14
Analysis based on clustering techniques.

Clustering algorithms	Number of articles	Reference number
K-means	10	Kan et al. (2010), Bodea et al. (2010), Chellatamilan et al. (2011), Bresfelean et al. (2012), Jacob et al. (2015), Mayilvaganan and Kalpanadevi (2015), Angra and Ahuja (2017), Akram et al. (2019), Francis and Babu (2019), M.Á. et al. (2020)
Expectation Maximization	1	Trandafili et al. (2012)
K-modes	1	El Aissaoui et al. (2020)

Table 15
Analysis based of Association Rule techniques.

Association Rule algorithms	Number of articles	Reference number
Apriori Association Rule algorithm	3	Kan et al. (2010), Trandafili et al. (2012), Rojanavasu (2019)
FP-growth	1	Mayilvaganan and Kalpanadevi (2015)
Relational Association rules	1	Jung (2018)

et al. (2013b,a), Mashiloane and Mchunu (2013), Palazuelos et al. (2013), Pathan et al. (2014), Anh et al. (2014), Chen et al. (2014), Manhães et al. (2014), Natek and Zwilling (2014), Pruthi and Bhatia (2015), Jacob et al. (2015), Parmar et al. (2015), Guarín et al. (2015), Sorour et al. (2015), Ahadi et al. (2015), Sisovic et al. (2015), Bakaric et al. (2015), Jishan et al. (2015), Kaur et al. (2015), Agaoglu (2016), Sanchez-Santillan et al. (2016), Meedech et al. (2016), Stahovich and Lin (2016), Badr et al. (2016), Ahmed et al. (2016), Kassak et al. (2016), Ayub et al. (2017), Athani et al. (2017), Castro-Wunsch et al. (2017), Buenaño-Fernández et al. (2017), Figueira (2017), Kitanaka et al. (2017), Leppänen et al. (2017), Pise and Kulkarni (2017), Jung (2016), Srivastava et al. (2018), Patil et al. (2018), Kaunang and Rotikan (2018), Chitra and Agrawal (2019), Spatiotis et al. (2018), Niu et al. (2018), Rustia et al. (2018), Chanlekha and Niramitranon (2018), Kularbphettong (2017), Burgos et al. (2018), Miguéis et al. (2018), Rojanavasu (2019), Ketui et al. (2019), Al Breiki et al. (2019), Lagman et al. (2019), Amazona and Hernandez (2019), Altaf et al. (2019), Al Fanah and Ansari (2019), Francis and Babu (2019), Adekitan and Salau (2020), Tarmizi et al. (2019), Rawat and Malhan (2019), Zaffar et al. (2018), Vila et al. (2018), Almutairi et al. (2019), Kasthuriarachchi and Liyanage (2018), Ab Ghani et al. (2018), Ibrahim et al. (2018), Yousafzai et al. (2020), Adekitan and Salau (2019), Lottering et al. (2020), Utari et al. (2020), M.Á. et al. (2020), Mengash (2020), Rahman and Mahmud (2020), Islam and Mahmud (2020), Karthikeyan et al. (2020), Ajibade et al. (2018), and Injadat et al. (2020b), out of 142 research articles which is 65% of total research articles considered for review purpose while k-fold parameter is considered in 66 research article which is 52% of total research articles. Performance parameters True positive rate, False positive rate, Precision, Recall, Receiver Operating Characteristic, F-measures, Correctly Classified Instances, Incorrectly Classified Instances, and Root mean square error are employed in 20, 17, 44, 38, 18, 37, 14, 10, and 11 number of research articles to evaluate various classification algorithms. The parameters Notch difference, Lift Value, and Cosine Cofficient are least used parameters in deciding the performance of classification algorithms.

4.7.1. Analysis on basis of accuracy

Accuracy is mostly used parameter for evaluating various classification algorithms. Accuracy is the ratio of correctly predicted value to total number of observation. The accuracy formula is

Accuracy = (True Positive +True Negative)/ (True Positive + False Positive + False Negative +True Negative) where

- True Positive — Actual class to which particular observation belongs is 'yes' but the predicted value for that observation is 'yes'.
- False Positive — Actual class to which particular observation belongs is 'no' but the predicted value for that observation is 'yes'.
- False Negative — Actual class to which particular observation belongs is 'yes' but the predicted value for that observation is 'no'.
- True Negative — Actual class to which particular observation belongs is 'no' but the predicted value for that observation is 'no'.

Analysis of research articles in terms of accuracy is illustrated in Table 17 with nine ranges <60%, 61%–65%, 66%–70%, 71%–75%, 76%–80%, 81%–85%, 86%–90%, 91%–95%, and 96%–100%. Thirteen research articles Sen and Ucar (2012), Márquez-Vera et al. (2013b), Natek and Zwilling (2014), Figueira (2017), Srivastava et al. (2018), Burgos et al. (2018), Miguéis et al. (2018), Tarmizi et al. (2019), Vila et al. (2018), Kasthuriarachchi and Liyanage (2018), Yousafzai et al. (2020), Lottering et al. (2020), Karthikeyan et al. (2020) showed the improved accuracy in the range 19%–100% while accuracy in the range 91%–95% is observed in the research articles Şen et al. (2012), Chau and Phung (2013), Márquez-Vera et al. (2013a), Jacob et al. (2015), Agaoglu (2016), Niu et al. (2018), Kularbphettong (2017), Ketui et al. (2019), Amazona and Hernandez (2019), Rawat and Malhan (2019), Utari et al. (2020), Ajibade et al. (2018).

The research articles Guarín et al. (2015), Francis and Babu (2019), Rahman and Mahmud (2020), Pruthi and Bhatia (2015), Kaunang and Rotikan (2018), Spatiotis et al. (2018), Islam and Mahmud (2020), Hoe et al. (2013), Palazuelos et al. (2013), Jishan et al. (2015), Badr et al. (2016), Castro-Wunsch et al. (2017), Dejaeger et al. (2012), Sorour et al. (2015), Ahadi et al. (2015), Kaur et al. (2015), Buenaño-Fernández et al. (2017), Patil et al. (2018), Rustia et al. (2018), Rojanavasu (2019), Almutairi et al. (2019), Ab Ghani et al. (2018), Ibrahim et al. (2018), and Pratiwi (2013), Sanchez-Santillan et al. (2016), Meedech et al. (2016), Kassak et al. (2016), Ayub et al. (2017), Kitanaka et al. (2017), Chitra and Agrawal (2019), Al Fanah and Ansari (2019), Adekitan and Salau (2020), Mengash (2020) noted the accuracy value in the range <60%, 61%–65%, 66%–70%, 71%–75%, and 76%–80% respectively.

The range 81%–85% and 86%–90% is noticed in the research articles Göker et al. (2013), Sisovic et al. (2015), Bakaric et al. (2015), Ahmed et al. (2016), Jung (2016), Al Breiki et al. (2019), Lagman et al. (2019), Altaf et al. (2019), M.Á. et al. (2020), and Cocea and Weibelzahl (2010), Nasiri et al. (2012), Mashiloane and Mchunu (2013), Pathan et al. (2014), Manhães et al. (2014), Stahovich and Lin (2016), Athani et al. (2017), Adekitan and Salau (2019).

4.7.2. Analysis on basis of TP and FP rate

This subsection discusses about the analysis of research articles in terms of TP rate and FP rate. Fig. 9 shows the analysis of TP rate with the five ranges 76%–80%, 81%–85%, 86%–90%, 91%–95% and 96%–100%. From Fig. 9, it is found that the research articles Chellatamilan et al. (2011), Márquez-Vera et al. (2013b), Gera and Goel (2015), and Vila et al. (2018) attained the improved TP rate in the range 96%–100% while the TP rate range in 91%–95% is observed in the research articles Márquez-Vera et al. (2013a), Ragab et al. (2014), Dimić et al. (2019), and Ashraf et al. (2020).

The research articles Manhães et al. (2014), Barbosa Manhães et al. (2015), Salinas and Stephens (2015), Auddy and Mukhopadhyay (2015), Chaudhury et al. (2016), Bodea et al. (2010), Cocea and Weibelzahl (2010), Kaur et al. (2015), and Bunkar et al. (2012),

Table 16
Analysis based on classification performance metric.

Performance Metric	No. of articles	% of No. of articles	Reference number
Accuracy	84	65%	Cocea and Weibelzahl (2010), Nasiri et al. (2012), Dejaeger et al. (2012), Şen et al. (2012), Sen and Ucar (2012), Hoe et al. (2013), Chau and Phung (2013), Pratiwi (2013), Göker et al. (2013), Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Mashiloane and Mchunu (2013), Palazuelos et al. (2013), Pathan et al. (2014), Anh et al. (2014), Chen et al. (2014), Manhães et al. (2014), Natek and Zwilling (2014), Pruthi and Bhatia (2015), Jacob et al. (2015), Parmar et al. (2015), Guarín et al. (2015), Sorour et al. (2015), Ahadi et al. (2015), Sisovic et al. (2015), Bakaric et al. (2015), Jishan et al. (2015), Kaur et al. (2015), Agaoglu (2016), Sanchez-Santillan et al. (2016), Meedech et al. (2016), Stahovich and Lin (2016), Badr et al. (2016), Ahmed et al. (2016), Kassak et al. (2016), Ayub et al. (2017), Athani et al. (2017), Castro-Wunsch et al. (2017), Buenaño-Fernández et al. (2017), Figueira (2017), Kitanaka et al. (2017), Leppänen et al. (2017), Pise and Kulkarni (2017), Jung (2016), Srivastava et al. (2018), Patil et al. (2018), Kaunang and Rotikan (2018), Chitra and Agrawal (2019), Spatiotis et al. (2018), Niu et al. (2018), Rustia et al. (2018), Chanlekha and Niramitranon (2018), Kularbphettong (2017), Burgos et al. (2018), Miguéis et al. (2018), Rojanavasu (2019), Ketui et al. (2019), Al Breiki et al. (2019), Lagman et al. (2019), Amazona and Hernandez (2019), Altaf et al. (2019), Al Fanah and Ansari (2019), Francis and Babu (2019), Adekitan and Salau (2020), Tarmizi et al. (2019), Rawat and Malhan (2019), Zaffar et al. (2018), Vila et al. (2018), Almutairi et al. (2019), Kasthuriarachchi and Liyanage (2018), Ab Ghani et al. (2018), Ibrahim et al. (2018), Yousafzai et al. (2020), Adekitan and Salau (2019), Lottering et al. (2020), Utari et al. (2020), M.Á. et al. (2020), Mengash (2020), Rahman and Mahmud (2020), Islam and Mahmud (2020), Karthikeyan et al. (2020), Ajibade et al. (2018), Injadat et al. (2020b)
True positive (TP) rate	20	16%	Bodea et al. (2010), Chellatamilan et al. (2011), Cocea and Weibelzahl (2010), Bunkar et al. (2012), Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Ragab et al. (2014), Manhães et al. (2014), Gera and Goel (2015), Barbosa Manhães et al. (2015), Salinas and Stephens (2015), Auddy and Mukhopadhyay (2015), Kaur et al. (2015), Mayilvaganan and Kalpanadevi (2015), Chaudhury et al. (2016), Martínez-Abad et al. (2018), Crivei et al. (2019), Dimić et al. (2019), Vila et al. (2018), Ashraf et al. (2020)
False positive (FP) rate	17	13%	Bodea et al. (2010), Chellatamilan et al. (2011), Cocea and Weibelzahl (2010), Bunkar et al. (2012), Mashiloane and Mchunu (2013), Ragab et al. (2014), Manhães et al. (2014), Gera and Goel (2015), Barbosa Manhães et al. (2015), Auddy and Mukhopadhyay (2015), Kaur et al. (2015), Mayilvaganan and Kalpanadevi (2015), Chaudhury et al. (2016), Crivei et al. (2019), Dimić et al. (2019), Injadat et al. (2020a), Ashraf et al. (2020)
False Negatives	2	2%	Manhães et al. (2014), Crivei et al. (2019)
True Negatives	4	3%	Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Manhães et al. (2014), Crivei et al. (2019)
Precision	44	34%	Bodea et al. (2010), Chellatamilan et al. (2011), Cocea and Weibelzahl (2010), El-Halees (2011), Bunkar et al. (2012), Göker et al. (2013), Mashiloane and Mchunu (2013), Chen et al. (2014), Ragab et al. (2014), Gera and Goel (2015), Guo et al. (2015), Sorour et al. (2015), Barbosa Manhães et al. (2015), Jishan et al. (2015), Auddy and Mukhopadhyay (2015), Kaur et al. (2015), Mayilvaganan and Kalpanadevi (2015), Amornsinlaphachai (2016), Agaoglu (2016), Chaudhury et al. (2016), Kassak et al. (2016), Kitanaka et al. (2017), Kiu (2018), Kaunang and Rotikan (2018), Lagus et al. (2018), Chanlekha and Niramitranon (2018), Martínez-Abad et al. (2018), Kularbphettong (2017), Burgos et al. (2018), Tasnim et al. (2019), Alshehri (2019), Francis and Babu (2019), Dimić et al. (2019), Vila et al. (2018), Almutairi et al. (2019), Ibrahim et al. (2018), Lottering et al. (2020), Utari et al. (2020), Mengash (2020), Injadat et al. (2020a), Karthikeyan et al. (2020), Ajibade et al. (2018), Ashraf et al. (2020), Injadat et al. (2020b)
Recall	38	30%	Bodea et al. (2010), Chellatamilan et al. (2011), El-Halees (2011), Bunkar et al. (2012), Göker et al. (2013), Chen et al. (2014), Ragab et al. (2014), Gera and Goel (2015), Sorour et al. (2015), Barbosa Manhães et al. (2015), Jishan et al. (2015), Auddy and Mukhopadhyay (2015), Kaur et al. (2015), Mayilvaganan and Kalpanadevi (2015), Amornsinlaphachai (2016), Agaoglu (2016), Kitanaka et al. (2017), Kiu (2018), Kaunang and Rotikan (2018), Lagus et al. (2018), Chanlekha and Niramitranon (2018), Kularbphettong (2017), Burgos et al. (2018), Tasnim et al. (2019), Alshehri (2019), Altaf et al. (2019), Francis and Babu (2019), Vila et al. (2018), Almutairi et al. (2019), Ab Ghani et al. (2018), Ibrahim et al. (2018), Lottering et al. (2020), Utari et al. (2020), Mengash (2020), Injadat et al. (2020a), Karthikeyan et al. (2020), Ajibade et al. (2018), Ashraf et al. (2020)
Confusion Matrix	3	2%	Barbosa Manhães et al. (2015), Agrawal et al. (2020), Ribeiro and Canedo (2020)
Area under Curve (AUC)	9	7%	Dejaeger et al. (2012), Jishan et al. (2015), Salinas and Stephens (2015), Niu et al. (2018), Rustia et al. (2018), Pérez et al. (2018), Crivei et al. (2019), Utari et al. (2020), Santoso (2019)
Receiver Operating Characteristic (ROC)	18	14%	Bodea et al. (2010), Chellatamilan et al. (2011), Bunkar et al. (2012), Chau and Phung (2013), Göker et al. (2013), Anh et al. (2014), Ragab et al. (2014), Gera and Goel (2015), Auddy and Mukhopadhyay (2015), Kaur et al. (2015), Mayilvaganan and Kalpanadevi (2015), Lehr et al. (2016), Chaudhury et al. (2016), Martínez-Abad et al. (2018), Chango et al. (2019), Vila et al. (2018), Ribeiro and Canedo (2020), Ashraf et al. (2020)
F-measures	37	29%	Bodea et al. (2010), Chellatamilan et al. (2011), El-Halees (2011), Bunkar et al. (2012), Göker et al. (2013), Chen et al. (2014), Ragab et al. (2014), Gera and Goel (2015), Sorour et al. (2015), Jishan et al. (2015), Auddy and Mukhopadhyay (2015), Kaur et al. (2015), Mayilvaganan and Kalpanadevi (2015), Amornsinlaphachai (2016), Kitanaka et al. (2017), Daud et al. (2017), Costa et al. (2017), Kiu (2018), Kaunang and Rotikan (2018), Lagus et al. (2018), Chanlekha and Niramitranon (2018), Kularbphettong (2017), Tasnim et al. (2019), Alshehri (2019), Chango et al. (2019), Francis and Babu (2019), Vila et al. (2018), Almutairi et al. (2019), Ibrahim et al. (2018), Lottering et al. (2020), Utari et al. (2020), Mengash (2020), Injadat et al. (2020a), Karthikeyan et al. (2020), Ajibade et al. (2018), Ashraf et al. (2020), Injadat et al. (2020b)
Correctly Classified Instances	14	11%	Bodea et al. (2010), Chellatamilan et al. (2011), Bresfelean et al. (2012), Pratiwi (2013), Göker et al. (2013), Mashiloane and Mchunu (2013), Dangi and Srivastava (2014), Barbosa Manhães et al. (2015), Mayilvaganan and Kalpanadevi (2015), Abyaa et al. (2018), Al Breiki et al. (2019), Chango et al. (2019), El Aissaoui et al. (2020), Ashraf et al. (2020)
Incorrectly Classified Instances	10	8%	Bodea et al. (2010), Chellatamilan et al. (2011), Pratiwi (2013), Göker et al. (2013), Mashiloane and Mchunu (2013), Dangi and Srivastava (2014), Barbosa Manhães et al. (2015), Mayilvaganan and Kalpanadevi (2015), El Aissaoui et al. (2020), Ashraf et al. (2020)

(*continued on next page*)

Table 16 (*continued*).

Performance Metric	No. of articles	% of No. of articles	Reference number
Kappa Statistic	6	5%	Bodea et al. (2010), Chellatamilan et al. (2011), Jacob et al. (2015), Martínez-Abad et al. (2018), Akram et al. (2019), El Aissaoui et al. (2020)
Cross-validation K-fold	66	52%	Chellatamilan et al. (2011), El-Halees (2011), Bresfelean et al. (2012), Şen et al. (2012), Sen and Ucar (2012), Chau and Phung (2013), Pratiwi (2013), Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Palazuelos et al. (2013), Anh et al. (2014), Ragab et al. (2014), Manhães et al. (2014), Guarín et al. (2015), Sorour et al. (2015), Ahadi et al. (2015), Sisovic et al. (2015), Bakaric et al. (2015), Auddy and Mukhopadhyay (2015), Amornsinlaphachai (2016), Lehr et al. (2016), Sanchez-Santillan et al. (2016), Meedech et al. (2016), Stahovich and Lin (2016), Ahmed et al. (2016), Ayub et al. (2017), Athani et al. (2017), Daud et al. (2017), Leppänen et al. (2017), Pise and Kulkarni (2017), Kaunang and Rotikan (2018), Lagus et al. (2018), Spatiotis et al. (2018), Niu et al. (2018), Rustia et al. (2018), Abyaa et al. (2018), Chanlekha and Niramitranon (2018), Martínez-Abad et al. (2018), Pérez et al. (2018), Burgos et al. (2018), Miguéis et al. (2018), Rojanavasu (2019), Tasnim et al. (2019), Ketui et al. (2019), Al Breiki et al. (2019), Akram et al. (2019), Alshehri (2019), Umer et al. (2019), Amazona and Hernandez (2019), Chango et al. (2019), Al Fanah and Ansari (2019), Martins et al. (2019), Rawat and Malhan (2019), Vila et al. (2018), Almutairi et al. (2019), Kasthuriarachchi and Liyanage (2018), Ab Ghani et al. (2018), Yousafzai et al. (2020), Utari et al. (2020), Mengash (2020), Rahman and Mahmud (2020), Santoso (2019), Ajibade et al. (2018), El Aissaoui et al. (2020), Ashraf et al. (2020), Injadat et al. (2020b)
Mean Absolute Error(MAE)	9	7%	Bodea et al. (2010), Chellatamilan et al. (2011), Dangi and Srivastava (2014), Gera and Goel (2015), Pruthi and Bhatia (2015), Amornsinlaphachai (2016), Abyaa et al. (2018), Al Breiki et al. (2019), El Aissaoui et al. (2020)
Root mean square error	11	9%	Bodea et al. (2010), Chellatamilan et al. (2011), Chuan et al. (2011), Wang and Liao (2011), Dangi and Srivastava (2014), Gera and Goel (2015), Al Breiki et al. (2019), Akram et al. (2019), Martins et al. (2019), Yousafzai et al. (2020), El Aissaoui et al. (2020)
Relative absolute error (RAE)	7	5%	Bodea et al. (2010), Chellatamilan et al. (2011), Dangi and Srivastava (2014), Gera and Goel (2015), Al Breiki et al. (2019), El Aissaoui et al. (2020), Ashraf et al. (2020)
Root relative squared error	7	5%	Bodea et al. (2010), Chellatamilan et al. (2011), Dangi and Srivastava (2014), Gera and Goel (2015), Martínez-Abad et al. (2018), Al Breiki et al. (2019), El Aissaoui et al. (2020)
Notch difference	1	1%	Dejaeger et al. (2012)
Lift Value	1	1%	Guruler et al. (2010)
Standard Deviation	3	2%	Meedech et al. (2016), Hassan and Al-Razgan (2016), Umer et al. (2019)
t-Test	2	2%	Zengin et al. (2011), Shukor et al. (2015)
Cosine Cofficient	1	1%	Sakurai et al. (2012)

Table 17
Analysis based on number of articles used accuracy as performance metric.

Range	Number of articles used accuracy as performance metric	Reference number
<60%	3	Guarín et al. (2015), Francis and Babu (2019), Rahman and Mahmud (2020)
61%–65%	4	Pruthi and Bhatia (2015), Kaunang and Rotikan (2018), Spatiotis et al. (2018), Islam and Mahmud (2020)
66%–70%	5	Hoe et al. (2013), Palazuelos et al. (2013), Jishan et al. (2015), Badr et al. (2016), Castro-Wunsch et al. (2017)
71%–75%	11	Dejaeger et al. (2012), Sorour et al. (2015), Ahadi et al. (2015), Kaur et al. (2015), Buenaño-Fernández et al. (2017), Patil et al. (2018), Rustia et al. (2018), Rojanavasu (2019), Almutairi et al. (2019), Ab Ghani et al. (2018), Ibrahim et al. (2018)
76%–80%	10	Pratiwi (2013), Sanchez-Santillan et al. (2016), Meedech et al. (2016), Kassak et al. (2016), Ayub et al. (2017), Kitanaka et al. (2017), Chitra and Agrawal (2019), Al Fanah and Ansari (2019), Adekitan and Salau (2020), Mengash (2020)
81%–85%	9	Göker et al. (2013), Sisovic et al. (2015), Bakaric et al. (2015), Ahmed et al. (2016), Jung (2016), Al Breiki et al. (2019), Lagman et al. (2019), Altaf et al. (2019), M.Á. et al. (2020)
86%–90%	8	Cocea and Weibelzahl (2010), Nasiri et al. (2012), Mashiloane and Mchunu (2013), Pathan et al. (2014), Manhães et al. (2014), Stahovich and Lin (2016), Athani et al. (2017), Adekitan and Salau (2019)
91%–95%	12	Şen et al. (2012), Chau and Phung (2013), Márquez-Vera et al. (2013a), Jacob et al. (2015), Agaoglu (2016), Niu et al. (2018), Kularbphettong (2017), Ketui et al. (2019), Amazona and Hernandez (2019), Rawat and Malhan (2019), Utari et al. (2020), Ajibade et al. (2018)
96%–100%	13	Sen and Ucar (2012), Márquez-Vera et al. (2013b), Natek and Zwilling (2014), Figueira (2017), Srivastava et al. (2018), Burgos et al. (2018), Miguéis et al. (2018), Tarmizi et al. (2019), Vila et al. (2018), Kasthuriarachchi and Liyanage (2018), Yousafzai et al. (2020), Lottering et al. (2020), Karthikeyan et al. (2020)

Mayilvaganan and Kalpanadevi (2015) obtained the TP rate value in the range 76%–80%, 81%–85%, and 86%–90% respectively.

Fig. 9 shows the analysis of FP rate with the six ranges 0%–5%, 6%–10%, 11%–15%, 16%–20%, 21%–25%, and 26%–30%. It is noted from Fig. 10 that the research articles Ragab et al. (2014), Mayilvaganan and Kalpanadevi (2015), Dimić et al. (2019), and Ashraf et al. (2020) attained low value of FP rate in the range 0%–5%.

FP rate range in 6%–10% and 11%–15% is observed in the research articles Bodea et al. (2010), Cocea and Weibelzahl (2010), Manhães et al. (2014), Barbosa Manhães et al. (2015), Chaudhury et al. (2016), and Chellatamilan et al. (2011), Bunkar et al. (2012) respectively.

The research articles Auddy and Mukhopadhyay (2015), and Mashiloane and Mchunu (2013), Gera and Goel (2015), Kaur et al. (2015) attained the FP rate value in the range 21%–25% and 26%–30% respectively.

4.7.3. Analysis on basis of precision and recall

This subsection discusses the analysis of research articles in terms of performance parameters precision and recall. The formula for precision is

Precision = True Positives/ (True Positives + False Positives) = True Positives/ Total Predicted Positives

while the formula for recall is

Recall = True Positives/ (True Positives + False Negatives) = True Positives/ Total Actual Positive.

Fig. 11 shows the graph for the analysis of precision with seven ranges <70%, 71%–75%, 76%–80%, 81%–85%, 86%–90%, 91%–95%, and 96%–100%.

Improved precision value in the range 91%–95% and 96%–100% had observed in the research articles Agaoglu (2016), Kularbphettong (2017), Dimić et al. (2019), Utari et al. (2020), Ashraf et al. (2020),

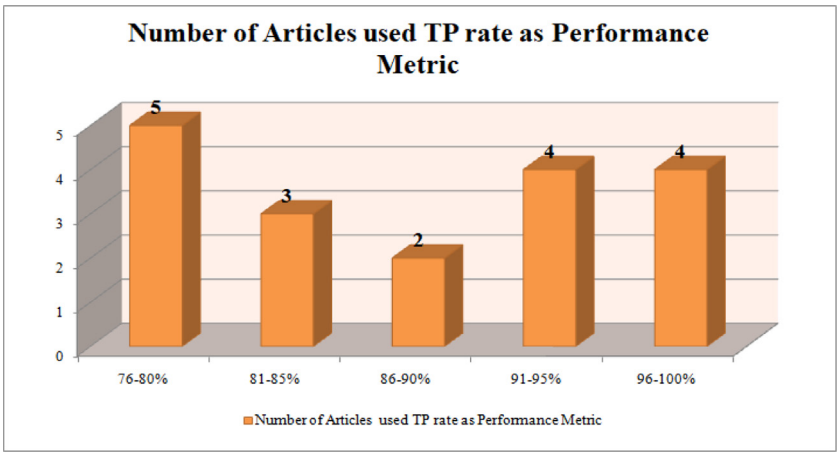


Fig. 9. Analysis on basis of performance metric TP rate.

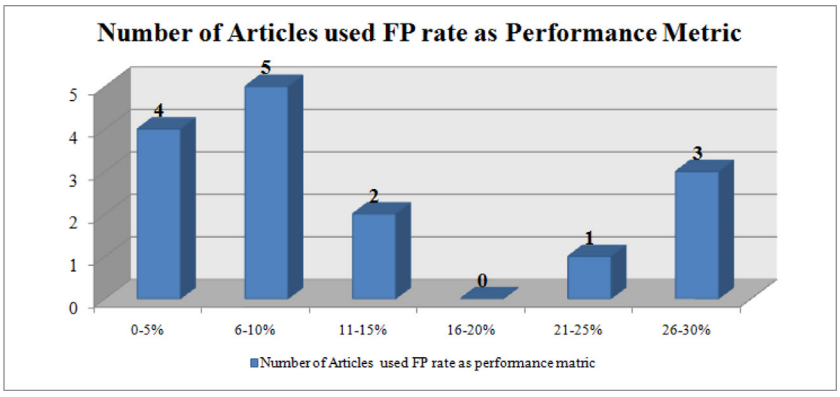


Fig. 10. Analysis on basis of performance metric FP rate.

and Burgos et al. (2018), Tasnim et al. (2019), Vila et al. (2018), Lottering et al. (2020), Karthikeyan et al. (2020) respectively.

While the research articles Gera and Goel (2015), Jishan et al. (2015), Kassak et al. (2016), Kaunang and Rotikan (2018), Martínez-Abad et al. (2018), Francis and Babu (2019), and Sorour et al. (2015), Amornsinlaphachai (2016), Chaudhury et al. (2016), Almutairi et al. (2019), Ibrahim et al. (2018) attained the comparatively low precision value in the range <70%, and 71%–75% respectively. In the research articles El-Halees (2011), Guo et al. (2015), Auddy and Mukhopadhyay (2015), Kitanaka et al. (2017), Mengash (2020), Bunkar et al. (2012), Göker et al. (2013), Barbosa Manhães et al. (2015), Kaur et al. (2015), and Mashiloane and Mchunu (2013), Ragab et al. (2014), Mayilvaganan and Kalpanadevi (2015), Kiu (2018), Ajibade et al. (2018), the precision value obtained are in the range 76%–80%, 81%–85%, and 86%–90% respectively.

Fig. 12 describe the analysis of recall value in the seven range <70%, 71%–75%, 76%–80%, 81%–85%, 86%–90%, 91%–95%, and 96%–100%.

Improved recall value in the range 91%–95% and 96%–100% had observed in the research articles Ragab et al. (2014), Agaoglu (2016), Ab Ghani et al. (2018), Karthikeyan et al. (2020), Ajibade et al. (2018), Ashraf et al. (2020), and Chellatamilan et al. (2011), Gera and Goel (2015), Kularbphettong (2017), Burgos et al. (2018), Tasnim et al. (2019), Vila et al. (2018), Lottering et al. (2020) respectively while the research articles Jishan et al. (2015), Kaunang and Rotikan (2018), Francis and Babu (2019), and Amornsinlaphachai (2016), Kitanaka et al. (2017), Almutairi et al. (2019), Ibrahim et al. (2018) attained the comparatively low recall value in the range <70%, and 71%–75% respectively. In the research articles El-Halees (2011), Sorour et al. (2015), Barbosa Manhães et al. (2015), Auddy and Mukhopadhyay

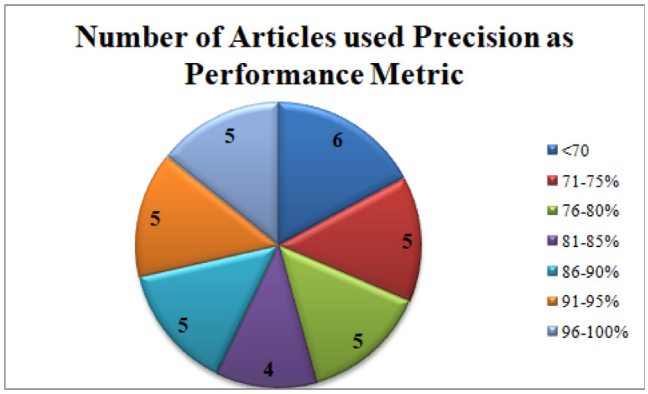


Fig. 11. Analysis on basis of performance metric Precision.

(2015), Bodea et al. (2010), Göker et al. (2013), Kaur et al. (2015), Altaf et al. (2019), Mengash (2020), and Kiu (2018), Utari et al. (2020), the recall value observed are in the range 76%–80%, 81%–85%, and 86%–90% respectively.

4.7.4. Analysis on basis of F-measures and cross validation

Analysis of research articles in terms of performance parameters F-measures and cross validation method – k-fold values is illustrated in this subsection. The formula for F-measures is

F-Measure = (2 * Precision * Recall)/ (Precision + Recall)

where Precision = True Positives/ Total Predicted Positives and Recall = True Positives/ Total Actual Positive.

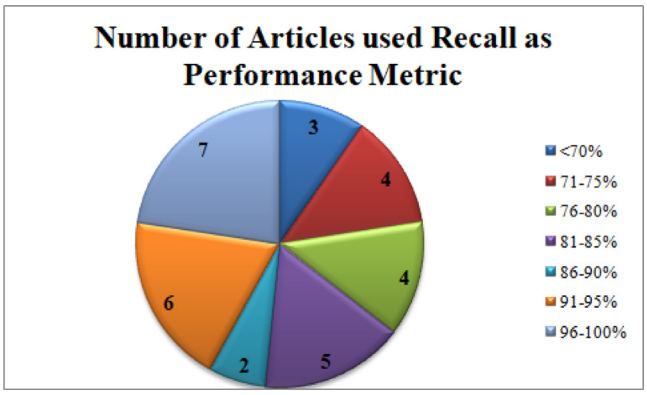


Fig. 12. Analysis on basis of performance metric Recall.

Fig. 13 shows the analysis of research articles which uses F-measure as performance metric with seven range values <70%, 71%–75%, 76%–80%, 81%–85%, 86%–90%, 91%–95%, and 96%–100%. The research articles Chellatamilan et al. (2011), Costa et al. (2017), Kularbphettong (2017), Ajibade et al. (2018), Ashraf et al. (2020), and Tasnim et al. (2019), Vila et al. (2018), Lottering et al. (2020), Karthikeyan et al. (2020) attained the improved value of f-measures in the range 91%–95% and 96%–100% respectively while low value of F-measure in the range <70% and 71%–75% had attained in the research articles Gera and Goel (2015), Jishan et al. (2015), Kaunang and Rotikan (2018), Francis and Babu (2019), and Bodea et al. (2010), Sorour et al. (2015), Amornsinlaphachai (2016), Kitanaka et al. (2017), Almutairi et al. (2019), Ibrahim et al. (2018) respectively. F-measure value in the range 76-8-%, 81%–85%, and 86%–90% is observed in the research articles El-Halees (2011), Auddy and Mukhopadhyay (2015), Mengash (2020), Bunkar et al. (2012), Göker et al. (2013), Kaur et al. (2015), Mayilvaganan and Kalpanadevi (2015), Chango et al. (2019), and Ragab et al. (2014), Daud et al. (2017), Kiu (2018), Utari et al. (2020) respectively.

Fig. 14 describes the analysis of research articles in terms of cross validation method with the k-fold values 2, 3, 4, 5, and 10. K-fold value is the validation technique in which dataset is divided into k-subsets. From Fig. 14, it is observed that 56 research articles used k-fold value that is 10 while seven research articles Anh et al. (2014), Ahmed et al. (2016), Daud et al. (2017), Kaunang and Rotikan (2018), Pérez et al. (2018), Rojanavasu (2019), Tasnim et al. (2019) considered k-fold value five.

4.7.5. Analysis on basis of correctly and incorrectly classified instance

Correctly and incorrectly classified instance performance parameters are also used to analyze the research articles. Fig. 15 shows the analysis of research articles based on correctly classified instance with five ranges 76%–80%, 81%–85%, 86%–90%, 91%–95% and 96%–100% while Fig. 16 describes the analysis of research articles based on incorrectly classified instance with four ranges 0%–5%, 6%–10%, 11%–15%, and 16%–20%.

The research articles Chellatamilan et al. (2011), Mayilvaganan and Kalpanadevi (2015), El Aissaoui et al. (2020), Al Breiki et al. (2019), Ashraf et al. (2020) had attained the improved correctly classified instance in the range 91%–95% and 96%–100% respectively while correctly classified instance in the ranges 76%–80%, 81%–85%, and 86%–90% are observed in the research articles Bresfelean et al. (2012), Pratiwi (2013), Bodea et al. (2010), Göker et al. (2013), Dangi and Srivastava (2014), Chango et al. (2019), and Dangi and Srivastava (2014), Barbosa Manhães et al. (2015) respectively.

Incorrectly classified instance in the range 0%–5% and 6%–10% attained in the research articles Mayilvaganan and Kalpanadevi (2015), Ashraf et al. (2020), and Chellatamilan et al. (2011), El Aissaoui et al. (2020) respectively. It is also noted from Fig. 16 that the research articles Göker et al. (2013), Mashiloane and Mchunu (2013), Dangi and Srivastava (2014), Barbosa Manhães et al. (2015), and Bodea et al. (2010), Pratiwi (2013) had the incorrectly classified instance in the range 11%–15% and 16%–20% respectively.

4.7.6. Analysis on basis of RMSE

This subsection illustrates the analysis in terms of Root Mean Square Error (RMSE) performance metric. RMSE is used to check how many concentrated data is around the best fit line. Table 18 illustrates the analysis based on number of articles used this performance parameter with range <0.1, 0.2–0.3, and >0.3. The range value less than <0.1 is observed in the research articles Wang and Liao (2011), Al Breiki et al. (2019), Akram et al. (2019), and El Aissaoui et al. (2020) while the research articles Dangi and Srivastava (2014), Martins et al. (2019), and Rawat and Malhan (2019) had attained the RMSE >3.0. The RMSE value in the range 0.2–0.3 is observed in the research articles Bodea et al. (2010), Chellatamilan et al. (2011), Chuan et al. (2011), and Gera and Goel (2015).

4.7.7. Analysis on basis of MAE

Analysis in terms of Mean Absolute Error (MAE) is discussed in this subsection. This MAE is used to measures the accuracy for continuous variables. Table 19 elucidates the analysis of research articles based

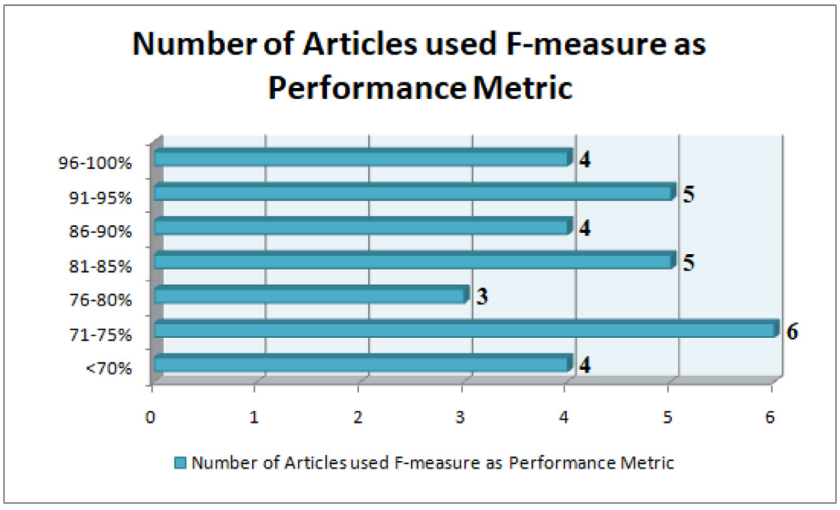


Fig. 13. Analysis on basis of performance metric F-measures.

Table 18
Analysis based on number of articles used RMSE as performance metric.

Range	Number of articles used RMSE as performance metric	Reference number
< 0.1	4	Wang and Liao (2011), Al Breiki et al. (2019), Akram et al. (2019), El Aissaoui et al. (2020)
0.2–0.3	4	Bodea et al. (2010), Chellatamilan et al. (2011), Chuan et al. (2011), Gera and Goel (2015)
>0.3	3	Dangi and Srivastava (2014), Martins et al. (2019), Rawat and Malhan (2019)

Table 19
Analysis based on number of articles used mean absolute error as performance metric.

Range	Number of articles used MAE as performance metric	Reference number
< 0.5	2	Pruthi and Bhatia (2015), El Aissaoui et al. (2020)
0.5–1.0	3	Chellatamilan et al. (2011), Gera and Goel (2015), Al Breiki et al. (2019)
1.1–1.5	1	Bodea et al. (2010)
>2.1	2	Dangi and Srivastava (2014), Amornsinlaphachai (2016)

on MAE performance metric with the range value <0.5, 0.5–1.0, 1.1–1.5, and >2.1. Two research articles Dangi and Srivastava (2014), Amornsinlaphachai (2016) obtained the MAE value >2.1 while <0.5 MAE value is observed in the research articles Pruthi and Bhatia (2015), El Aissaoui et al. (2020).

4.7.8. Analysis on basis of ROC

ROC is used to show the performance of classification model at all classification thresholds. To draw the ROC graph, two parameters — True Positive Rate and False Positive rates are used. Low value of classification threshold indicate more positive items. Analysis in terms of receiver operating characteristic (ROC) curve is discussed in this subsection.

Table 21 illustrates the analysis on basis of ROC performance metric with five range values <80%, 81%–85%, 86%–90%, 91%–95%, and 96%–100%. It is noted from Table 21 that the research articles Bunkar et al. (2012), Chango et al. (2019), Bodea et al. (2010), Chellatamilan et al. (2011), Chau and Phung (2013), Ragab et al. (2014), Mayilvaganan and Kalpanadevi (2015), and Ashraf et al. (2020) had attained the improved value of ROC in the range 96%–100% while < 80% range value of ROC attained in the research articles Auddy and Mukhopadhyay (2015), Kaur et al. (2015), Chaudhury et al. (2016), Martínez-Abad et al. (2018).

The ROC value in the range 81%–85%, 86%–90%, and 91%–95% is observed in the research articles Gera and Goel (2015), Lehr et al. (2016), Vila et al. (2018), Göker et al. (2013), Ribeiro and Canedo (2020), and Bunkar et al. (2012), Chango et al. (2019) respectively.

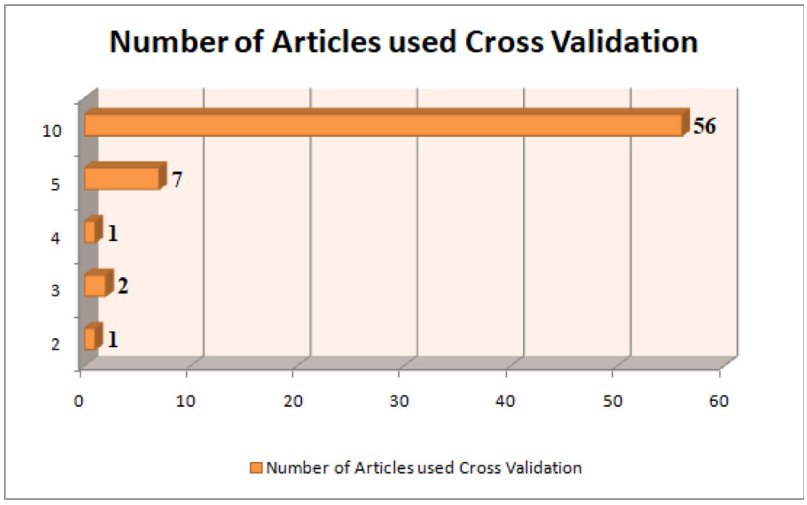


Fig. 14. Analysis on basis of performance metric Cross Validation.

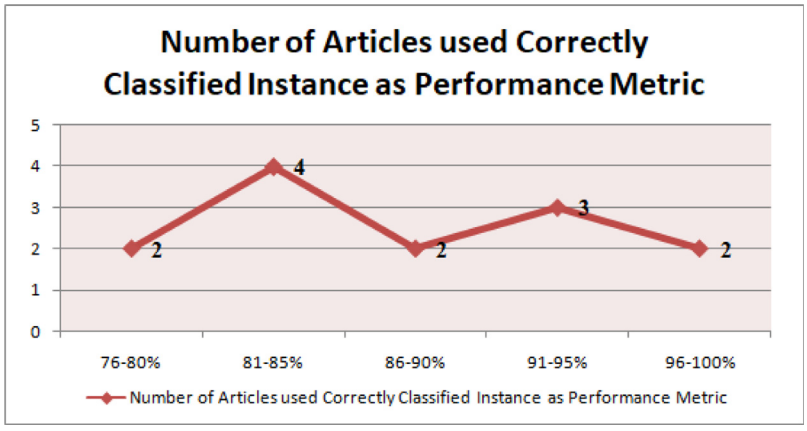


Fig. 15. Analysis on basis of performance metric correctly classified instance.

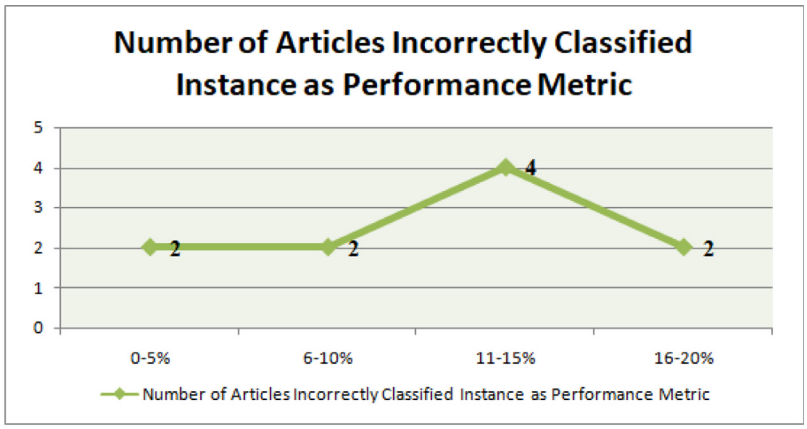


Fig. 16. Analysis on basis of performance metric incorrectly classified instance.

Table 20
Analysis based on number of articles used AUC as performance metric.

Range	Number of articles used AUC as performance metric	Reference number
<80%	2	Chitra and Agrawal (2019), Attiya et al. (2022)
86%–90%	2	Salinas and Stephens (2015), Santoso (2019)
91%–95%	3	Dejaeger et al. (2012), Niu et al. (2018), Pérez et al. (2018)
96%–100%	1	Utari et al. (2020)

Table 21
Analysis based on number of articles used ROC as performance metric.

Range	Number of articles used ROC as performance metric	Reference number
<80%	4	Auddy and Mukhopadhyay (2015), Kaur et al. (2015), Chaudhury et al. (2016), Martínez-Abad et al. (2018)
81%–85%	3	Gera and Goel (2015), Lehr et al. (2016), Vila et al. (2018)
86%–90%	2	Göker et al. (2013), Ribeiro and Canedo (2020)
91%–95%	2	Bunkar et al. (2012), Chango et al. (2019)
96%–100%	6	Bunkar et al. (2012), Chango et al. (2019) Bodea et al. (2010), Chellatamilan et al. (2011), Chau and Phung (2013), Ragab et al. (2014), Mayilvaganan and Kalpanadevi (2015), Ashraf et al. (2020)

4.7.9. Analysis on basis of AUC

Area under ROC (AUC) is used to measure total area underneath ROC curve from (0, 0) to (1, 1). It is used for binary classification problems. If the value of AUC is higher then the performance of model is better.This subsection illustrates the analysis of area under ROC (AUC).

Table 20 elucidates the analysis of research articles based on AUC performance metric in the four range values <80%, 86%–90%, 91%–95%, and 96%–100%. One research article Utari et al. (2020) attained the improved AUC value in the range 96%–100% while the AUC value in the range 91%–95% is observed in the research articles Dejaeger et al. (2012), Niu et al. (2018), Pérez et al. (2018).

Chitra and Agrawal (2019), Attiya et al. (2022), and Salinas and Stephens (2015), Santoso (2019) obtained the AUC value in the range <80% and 86%–90% respectively.

4.8. Analysis on the basis of selecting the best classification technique

Analysis in terms of selection of the best classification technique after comparing various classification techniques based various performance parameters is presented in this subsection.

Table 22 elucidates the analysis of research articles for selecting the best classification technique on given dataset along with classification techniques used and performance parameters. Naïve Bays is found to be the best classification technique in the research articles El-Halees (2011), Göker et al. (2013), Chen et al. (2014), Manhães et al. (2014), Guarín et al. (2015), Chaudhury et al. (2016), Kiu (2018), Santoso (2019), and Ashraf et al. (2020) while Random Forest is noted to be the best in the research articles Ragab et al. (2014), Ahadi et al. (2015), Lagus et al. (2018), Spatiotis et al. (2018), Pérez et al. (2018), Akram et al. (2019), Almutairi et al. (2019), and Adekitan and Salau (2019). The research articles Sorour et al. (2015), Bakaric et al. (2015), Daud et al. (2017), Costa et al. (2017), Tarmizi et al. (2019), Ibrahim et al. (2018), and Lottering et al. (2020) considered Support vector machine as the best classification technique while the research article Sen and Ucar (2012), Ragab et al. (2014), Sisovic et al. (2015), Amornsinlaphachai (2016), Agaoglu (2016), Kaunang and Rotikan (2018), Rustia et al. (2018), Kasthuriarachchi and Liyanage (2018), Ab Ghani et al. (2018), and Islam and Mahmud (2020) found Decision tree as the best classification algorithm.

The research articles Pise and Kulkarni (2017), Rahman and Mahmud (2020), Agrawal et al. (2020), Pratiwi (2013), Mashiloane and Mchunu (2013), Palazuelos et al. (2013), and Barbosa Manhães et al. (2015), Jishan et al. (2015), Kaur et al. (2015), Castro-Wunsch et al. (2017), Jung (2016), Chitra and Agrawal (2019) noted the best classification algorithm as K-nearest neighbour, J48, and Multilayer perceptron respectively.

4.9. Analysis on the basis of software/ Languages used in EDM

To implement the developed model using various data mining techniques in EDM, various softwares/languages such as Java, R-Programming, Python programming, etc. are used. Generally R-programming and Python programming are used to analyze the data and build the model in EDM. This subsection discusses the analysis based on the softwares/languages used such as ANSI C, Python, ASP.NET, C# programming, Java, MATLAB, R programming and SQL to build the model in EDM. Fig. 17 shows the analysis of research articles based on software used in building the model to predict students' performance in EDM.

The research articles Guo et al. (2015), Pérez et al. (2018), Umer et al. (2019), Al Fanah and Ansari (2019), M.Á. et al. (2020), and Maitra et al. (2018), Lottering et al. (2020), Jung (2018), Agrawal et al. (2020), Injadat et al. (2020b) used Python and R-programming to

Table 22
Analysis based on selecting the best classification technique.

Ref.	Article title	Classification techniques	Performance parameters	Best on dataset
Cocea and Weibelzahl (2010)	Disengagement Detection in Online Learning: Validation Studies and Perspectives	Bayesian Nets, Logistic regression, Simple logistic classification, Instance-based classification with IBk algorithm, Attribute Selected Classification using J48 classifier and Best First search, Bagging using REP tree classifier, Classification via Regression, Decision Trees with J48 classifier	Accuracy, True positive rate, False positive rate, Precision, Error	Simple logistic classification
Chuan et al. (2011)	Combining Different Classifiers in Educational Data Mining	Logistic Regression, K-nearest neighbour and SVD (Singular Value Decomposition)	Root mean square error	Logistic Regression
El-Halees (2011)	Mining Opinions in User-Generated Contents to Improve Course Evaluation	K-nearest neighbour,Naïve Bays, Support Vector Machine	Precision, Recall, F-measures, Cross-validation K-fold	Naïve Bays
Dejaeger et al. (2012)	Gaining insight into student satisfaction using comprehensible data mining techniques	Classification And Regression Trees, Multilayered perceptrons, Oblique Classifier 1, Logit, Support Vector Machine	Accuracy, Area under Curve, Notch difference	Logit
Sen and Ucar (2012)	Evaluating the achievements of computer engineering department of distance education students with data mining methods	Artificial neural networks, and Decision tree	Accuracy, Cross-validation K-fold	Decision tree
Chau and Phung (2013)	Imbalanced educational data classification: an effective approach with resampling and random forest	Naïve Bayes, Support Vector Machine, Neural Network, Logistic Regression, k-nearest Neighbour, and Decision Tree C4.5, Random Forest	Accuracy, Receiver Operating Characteristic, Cross-validation K-fold	Random Forest
Pratiwi (2013)	Predicting Student Placement Class using Data Mining	J48, SimpleCart, Kstar, Sequential minimal optimization, Naive Bayes, OneR	Accuracy, Correctly Classified Instances, Incorrectly Classified Instances, Cross-validation K-fold	J48
Göker et al. (2013)	The Estimation of Students' Academic Success by Data Mining Methods	Naive Bayes, J48, Bays Net, Radial basis function Network	Accuracy, Precision, Recall, Receiver Operating Characteristic, F-measures, Correctly Classified Instances, Incorrectly Classified Instances	Naive Bayes
Márquez-Vera et al. (2013b)	Predicting School Failure and Dropout by Using Data Mining Techniques	Jrip, Nnge, OneR, Prism, Ridor, J48, C4.5, SimpleCart, ADTree, Random Tree	Accuracy, True positive rate, True Negatives rate, Cross-validation K-fold	ADTree
Márquez-Vera et al. (2013a)	Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data	Jrip, Nnge, OneR, Prism, Ridor, J48, C4.5, SimpleCart, ADTree, Random Tree, REPTree, Interpretable Classification Rule Mining	Accuracy, True positive rate, True Negatives rate, Cross-validation K-fold	Interpretable Classification Rule Mining
Mashiloane and Mchunu (2013)	Mining for Marks: A Comparison of Classification Algorithms when Predicting Academic Performance to Identify "Students at Risk"	J48 classifier, Naïve Bayes and Decision Table	Accuracy, False positive rate, Precision, Correctly Classified Instances, Incorrectly Classified Instances	J48
Palazuelos et al. (2013)	Social Network Analysis and Data Mining: An Application to the E-Learning Context	J48, Random forests, Naïve Bayes, Bayesian networks, JRip, and Ridor	Accuracy, Cross-validation K-fold	J48
Chen et al. (2014)	Mining Social Media Data for Understanding Students' Learning Experiences	Naïve Bayes , MaxMargin Multi-Label classifier, Support vector machine	Accuracy, Precision, Recall, F-measures	Naïve Bayes
Ragab et al. (2014)	A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining	C4.5, Random Forest, IBK-E, IBK (Instance Based Learner)-M, LibSVM, MLP (Multilayer perceptron), Multilayer Perceptron, Naïve Bayes, and PART	True positive rate, False positive rate,Precision, Recall, Receiver Operating Characteristic, F-measures, Cross-validation K-fold	C4.5, PART and Random Forest
Manhães et al. (2014)	WAVE: an Architecture for Predicting Dropout in Undergraduate Courses using EDM	Naïve Bayes, Multilayer Perceptron,Support Vector Machine with polynomial kernel and RBF (Radial basis function kernel) kernel and Decision Table	Accuracy, True positive rate, False positive rate, False Negatives Rate, True Negatives rate, Cross-validation K-fold	Naïve Bayes

(*continued on next page*)

Table 22 (*continued*).

Ref.	Article title	Classification techniques	Performance parameters	Best on dataset
Natek and Zwilling (2014)	Student data mining solution–knowledge management system related to higher education institutions	J48, M5P and RepTree	Accuracy	RepTree
Guo et al. (2015)	Predicting Students Performance in Educational Data Mining	Naïve Bayes, Multilayer Perception and Support vector machine, SPPN Prediction Network	Precision	SPPN Prediction Network
Guarín et al. (2015)	A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining	Naïve Bayes and a decision tree	Accuracy, Cross-validation K-fold	Naïve Bayes
Sorour et al. (2015)	Correlation of Topic Model and Student Grades Using Comment Data Mining	Support Vector Machine and Artificial Neural Network	Accuracy, Precision, Recall, F-measures, Cross-validation K-fold	Support Vector Machine
Ahadi et al. (2015)	Exploring Machine Learning Methods to Automatically Identify Students in Need of Assistance	Naive Bayes, Bayesian Network, Decision Table, Conjuctive Rule, PART, ADTree, J48, Random Forest, Decision Stum	Accuracy, Cross-validation K-fold	Random Forest
Sisovic et al. (2015)	Mining Student Data to Assess the Impact of Moodle Activities and Prior Knowledge on Programming Course Success	C4.5, JRip and PART	Accuracy, Cross-validation K-fold	C4.5
Bakaric et al. (2015)	Text Mining Student Reports	C4.5, PART, NaiveBayes, and SVM (Support Vector Machine)	Accuracy, Cross-validation K-fold	Support Vector Machine (SVM)
Barbosa Manhães et al. (2015)	Towards Automatic Prediction of Student Performance in STEM Undergraduate Degree Programs	BayesNet, J48 ,JRip, Support Vector Machines, Multilayered Perceptrons, AdaBoost, SimpleLogistic, DecisionTable, OneR , RandomForest.	True positive rate, False positive rate, Precision, Recall, Confusion Matrix,Correctly Classified Instances , Incorrectly Classified Instances	Multilayered Perceptrons (MLP), Simple Logistic (SL), Random Forest (RF), SVM
Jishan et al. (2015)	Application of Optimum Binning Technique in Data Mining Approaches to Predict Students' Final Grade in a Course	Naive Bayes, C4.5, Neural Network	Accuracy, Precision, Recall, Area under Curve, F-measures	Neural Network
Auddy and Mukhopadhyay (2015)	Data Mining on ICT Usage in an Academic Campus: A Case Study	J48, JRip, QuickRules Fuzzy-Rough rule induction, Fuzzy Nearest Neighbour, Fuzzy Rough Nearest Neighbour and Vaguely Quantified Nearest Neighbour, Fuzzy-Rough Feature Selection	True positive rate, False positive rate, Precision, Recall, Receiver Operating Characteristic, F-measures, Cross-validation K-fold	Fuzzy Nearest Neighbour, Fuzzy Rough Nearest Neighbour
Kaur et al. (2015)	Classification and prediction based data mining algorithms to predict slow learners in education sector	Multilayer Perception, Naïve Bayes, Sequential minimal optimization, J48 and REPTree	Accuracy, True positive rate, False positive rate, Precision, Recall, Receiver Operating Characteristic, F-measures	Multilayer Perception
Amornsinlaphachai (2016)	Efficiency of data mining models to predict academic performance and a cooperative learning model	Artificial Neural Network, K-Nearest Neighbour, Naive Bayes, Bayesian Belief Network, JRIP, ID3 and C4.5	Precision, Recall, F-measures, Cross-validation K-fold, Mean Absolute Error	C4.5
Lehr et al. (2016)	Use Educational Data Mining to Predict Undergraduate Retention	Logistic Regression, Naïve Bayes, K Nearest Neighbourhood, Random Forest, Multilayer Perceptron, and Decision Tree	Receiver Operating Characteristic, Cross-validation K-fold	Logistic Regression
Agaoglu (2016)	Predicting Instructor Performance Using Data Mining Techniques in Higher Education	Decision tree algorithms — C5.0 and CART (Classification And Regression Trees) , Support vector machines, Artificial neural networks, and Discriminant analysis	Accuracy, Precision, Recall	C5.0
Chaudhury et al. (2016)	Enhancing the capabilities of Student Result Prediction System	Support vector machine, C4.5 and Naïve Bayes.	True positive rate, False positive rate, Precision, Receiver Operating Characteristic	Naïve Bayes
Meedech et al. (2016)	Prediction of Student Dropout Using Personal Profile and Data Mining Approach	JRip, OneR, Ridor, J48, SimpleCart, ADTree, RandomTree and REPTree	Accuracy, Cross-validation K-fold, Standard Deviation	ADTree

(*continued on next page*)

Table 22 (*continued*).

Ref.	Article title	Classification techniques	Performance parameters	Best on dataset
Ahmed et al. (2016)	Using data mining to predict instructor performance	J48 , Multilayer Perceptron, Naïve Bayes , Sequential Minimal Optimization	Accuracy, Cross-validation K-fold	Sequential minimal optimization and Multilayer Perceptron
Castro-Wunsch et al. (2017)	Evaluating Neural Networks as a Method for Identifying Students in Need of Assistance	NaiveBayes, Neural Network, RandomForest, DecisionTable, DecisionStump, PART, J48	Accuracy	Neural Network
Daud et al. (2017)	Predicting Student Performance using Advanced Learning Analytics	Support Vector Machine, C4.5, Classification and Regression Tree, Bayes Network, Naive Bayes	F-measures, Cross-validation K-fold	Support Vector Machine
Pise and Kulkarni (2017)	Evolving learners' behavior in data mining	k-nearest Neighbour, Naïve Bayes, IBK, J48, Adaboost, Logitboost, PART, Random Forest, Bagging, SMO	Accuracy, Cross-validation K-fold	k-nearest Neighbour
Jung (2016)	A Comparison of Data Mining Methods in Analyzing Educational Data	Neural Network, Multi-Layer Perceptron, Logistic Regression, Decision Tree -Chi-square Automatic Interaction Detector for Decision Tree	Accuracy	Neural Network - Multi-Layer Perceptron
Costa et al. (2017)	Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses	Neural Networks, Decision Tree, Support Vector Machine, Naive Bayes	F-measures	Support Vector Machine
Kiu (2018)	Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities	Naïve Bayesian, Multilayer Perceptron, Decision Tree J48 and Random Forest	Precision, Recall, F-measures	Na´ive Bayesian
Kaunang and Rotikan (2018)	Students' Academic Performance Prediction using Data Mining	Random Forest and Decision Tree	Accuracy, Precision, Recall, F-measures, Cross-validation K-fold	Decision Tree
Lagus et al. (2018)	Transfer-Learning Methods in Programming Course Outcome Prediction	Support Vector Machine, Random Forest, and AdaBoost	Precision, Recall, F-measures, Cross-validation K-fold	Random Forest
Spatiotis et al. (2018)	Evaluation of an Educational Training Platform Using Text Mining	REPTree, CART (Classification And Regression Trees), Random Forest, J48, Ibk, Bagging, AdaBoostM1, SVM, Neural Network	Accuracy, Cross-validation K-fold	Random Forest
Niu et al. (2018)	Exploring Causes for the Dropout on Massive Open Online Courses	Generalized Linear Models, Support Vector Machines, Tree Based Ensemble Learning Models, Neural Network Models, Naïve Bayes Models, K-Neighbours, Gaussian Progress.	Accuracy, Area under Curve, Cross-validation K-fold	XGBoost
Rustia et al. (2018)	Predicting Student's Board Examination Performance using Classification Algorithms	Naïve Bayes, Support Vector Machine, Neural Network, C4.5 Decision Tree, Logistic Regression	Accuracy, Area under Curve, Cross-validation K-fold	C4.5 Decision Tree
Ku-larbphet-tong (2017)	Analysis of Students' Behavior Based on Educational Data Mining	J48, Bayesian networks	Accuracy, Precision, Recall, F-measures	Bayesian networks
Pérez et al. (2018)	Predicting Student Drop-Out Rates Using Data Mining Techniques: A Case Study	Decision Trees, Logistic Regression, Naive Bayes and Random Forest	Area under Curve, Cross-validation K-fold	Random Forest
Chitra and Agrawal (2019)	Analysis of Educational Data Mining using Classification	J48, Support Vector machine, Naïve Bayes, Random Forest, Multilayer Perceptron	Accuracy	Multilayer Perceptron
Ketui et al. (2019)	Using Classification Data Mining Techniques for Students Performance Prediction	Decision Tree, Decision Tree Weight-based, ID3, Random Tree, and Gradient boosted tree	Accuracy, Cross-validation K-fold	Gradient boosted tree

(*continued on next page*)

Table 22 (*continued*).

Ref.	Article title	Classification techniques	Performance parameters	Best on dataset
Al Breiki et al. (2019)	Using Educational Data Mining Techniques to Predict Student Performance	Simple Log Regression, Decision Table, Gaussian Processes Random Tree, propositional rule learner, K-nearest neighbours, random forest, multi-layer perceptron, Naïve Bayes, Bayes Network learning, Multilayer Perceptron, Linear Regression, Random Forest, and Sequential Minimal Optimization	Accuracy, Correctly Classified Instances, Cross-validation K-fold, Mean Absolute Error, Root mean square error, Relative absolute error, Root relative squared error	Sequential Minimal Optimization
Akram et al. (2019)	Predicting Students' Academic Procrastination in Blended Learning Course Using Homework Submission Data	ZeroR, OneR, ID3, J48, Random forest, decision stump, JRip, PART, NBTree and Prism	Kappa Statistic, Cross-validation K-fold, Root mean square error	Random forest
Amazona and Hernandez (2019)	Modeling Student Performance Using Data Mining Techniques: Inputs for Academic Program Development	Naïve Bayes, Deep Learning in Neural Network, and Decision Tree	Accuracy, Cross-validation K-fold	Deep Learning in Neural Network
Chango et al. (2019)	Predicting academic performance of university students from multi-sources data in blended learning	J48, REPTREE, RamdomTree, JRIP, Nnge, PART	Receiver Operating Characteristic , F-measures, Correctly Classified Instances, Cross-validation K-fold	PART
Al Fanah and Ansari (2019)	Understanding E-learners' Behavior Using Data Mining Techniques	Random Forests, Logistic Regressions and Bayesian Networks	Accuracy, Cross-validation K-fold	Bayesian Networks
Tarmizi et al. (2019)	A Case Study on Student Attrition Prediction in Higher Education Using Data Mining Techniques	J48, Random Forest, Naïve Bayes, Support vector machine (RBF Kernal), Support vector machine (Polynomial Kernal)	Accuracy	Support vector machine (Polynomial Kernal)
Vila et al. (2018)	Detection of Desertion Patterns in University Students Using Data Mining Techniques: A Case Study	RandomTree, Naïve Bayes	Accuracy, Precision, Recall, Receiver Operating Characteristic, F-measures, Cross-validation K-fold	RandomTree
Almutairi et al. (2019)	Predicting Students' Academic Performance and Main Behavioral Features Using Data Mining Techniques	Random forest , Extreme gradient boosting, Logistic regression, Multilayer perceptron artificial neural network	Accuracy, Precision, Recall, F-measures, Cross-validation K-fold	Random Forest
Kasthuriarachchi and Liyanage (2018)	Predicting Students' Academic Performance Using Utility Based Educational Data Mining	Naïve Bayes, Random Forest and Decision Tree	Accuracy, Cross-validation K-fold	Decision Tree
Ab Ghani et al. (2018)	Student Enrollment Prediction Model in Higher Education Institution: A Data Mining Approach	Logistic regression, Decision tree, Naïve Bayes	Accuracy, Recall, Cross-validation K-fold	Decision tree
Ibrahim et al. (2018)	Mining Unit Feedback to Explore Students' Learning Experiences	Support Vector Machines, Naive Bayes, Decision Tree, Random Forest	Accuracy, Precision, Recall, F-measures	Support Vector Machines
Adekitan and Salau (2019)	The impact of engineering students' performance in the first three years on their graduation result using educational data mining	Probabilistic Neural Network based on the DDA (Dynamic Decay Adjustment), Random Forest Predictor, Decision Tree Predictor, Naïve Bayes Predictor, Tree Ensemble Predictor, and Logistic Regression Predictor	Accuracy	Random Forest
Lottering et al. (2020)	A model for the identification of students at risk of dropout at a university of technology	Decision Trees (e5.0), Support Vector Machine , Naive Bayes and Nearest Neighbour	Accuracy, Precision, Recall, F-measures	Support Vector Machine
Mengash (2020)	Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems	Artificial Neural Network, Decision Trees, Support Vector Machines, and Naïve Bayes	Accuracy, Precision, Recall, F-measures, Cross-validation K-fold	Artificial Neural Network

(*continued on next page*)

build the model while Java and SQL is considered in the research articles Parmar et al. (2015), Badr et al. (2016), Costa et al. (2017), Utari et al. (2020), and Guruler et al. (2010), Blagojević and Micić (2013), Gera and Goel (2015) respectively. The software such as ANSI C, ASP.NET, C# programming, and MATLAB is used in the research articles Guo et al. (2015), Gera and Goel (2015), Ramanathan et al.

Table 22 (*continued*).

Ref.	Article title	Classification techniques	Performance parameters	Best on dataset
Rahman and Mahmud (2020)	Classification on Educational Performance Evaluation Dataset using Feature Extraction Approach	K-Nearest Neighbours, Support Vector Machine and Convolutional Neural Network	Accuracy, Cross-validation K-fold	K-Nearest Neighbours
Islam and Mahmud (2020)	Integration of Learning Analytics into Learner Management System using Machine Learning	Logistic Regression (LR), k-nearest neighbours (KNN) and Decision Trees (DT)	Accuracy	Decision Trees
Santoso (2019)	The Analysis of Student Performance Using Data Mining	Decision tree (C4.5), and Naïve Bayes	Area under Curve, Cross-validation K-fold	Naïve Bayes
El Aissaoui et al. (2020)	Mining Learners' Behaviors: An Approach Based on Educational Data Mining Techniques	Naive Bayes, Classification and Regression Trees, ID3, and C4.5	Correctly Classified Instances, Incorrectly Classified Instances, Kappa Statistic, Cross-validation K-fold, Mean Absolute Error, Root mean square error, Relative absolute error, Root relative squared error	ID3
Agrawal et al. (2020)	Performance Appraisal of an Educational Institute Using Data Mining Techniques	Linear Discriminant Analysis, Classification and Regression Trees, K-Nearest Neighbour, Support Vector Machines, and Random Forest	Confusion Matrix, Receiver Operating Characteristic (ROC)	K-Nearest Neighbour
Ribeiro and Canedo (2020)	Using DataMining Techniques to Perform School Dropout Prediction: A Case Study	Generalized Linear Model, Gradient Boosting Machine, Support Vector Machines, Random Forest	True positive rate, False positive rate,Precision, Recall, Receiver Operating Characteristic, F-measures, Correctly Classified Instances, Incorrectly Classified Instances, Cross-validation K-fold, Relative absolute error	GBM: Gradient Boosting Machine
Ashraf et al. (2020)	An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering approaches	J48, Random tree, Naïve bayes, K-nearest neighbour, and Boosting	Accuracy, Precision, F-measures, Cross-validation K-fold	Naïve bayes

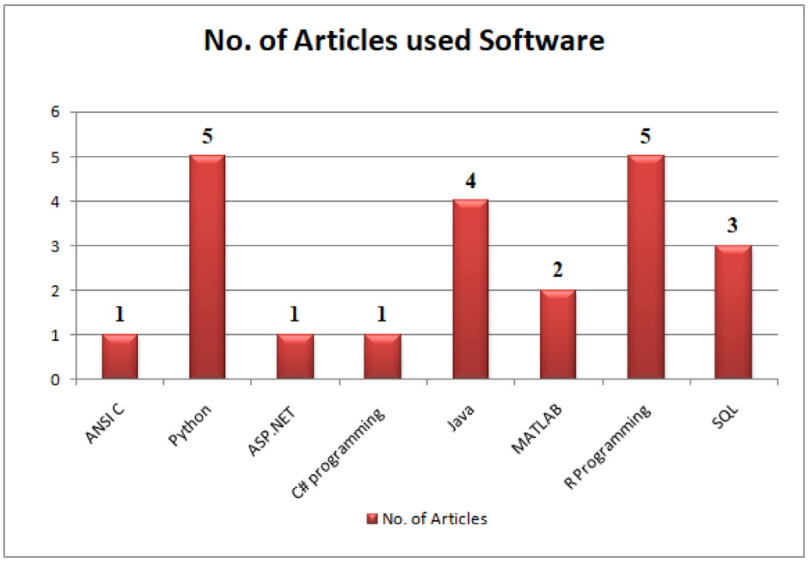


Fig. 17. Analysis on basis of software.

(2016), and Tasnim et al. (2019), Ajibade et al. (2018) respectively to predict students' performance in EDM.

4.10. Analysis on the basis of sampling period

Analysis in terms of sampling period considered for collecting the dataset is presented in this section with number of years <1 Year, 1 Year, 1 and $\frac{1}{2}$ Year,2 Years, 2 and $\frac{1}{2}$ Years, 3 Years, 4 Years, 5 Years, 6 Years, 7 Years, 9 Years, and 12 Years. Table 23 gives the analysis of research articles based on sampling period considered for collecting the data. Thirteen research articles Chuan et al. (2011), Dejaeger et al. (2012), Şen et al. (2012), Pratiwi (2013), Márquez-Vera et al. (2013b), Guarín et al. (2015), Sisovic et al. (2015), Lehr et al. (2016), Meedech et al. (2016), Costa et al. (2017), Kularbphettong (2017), Altaf et al. (2019), and Karthikeyan et al. (2020) considered the sampling period of 1 year for collecting students' data.

Sampling period of 2, 3, 4, and 5 years had been used in the research articles Mashiloane and Mchunu (2013), Sanchez-Santillan et al. (2016), Al Breiki et al. (2019), Martins et al. (2019), Chau and Phung (2013), Palazuelos et al. (2013), Natek and Zwilling (2014), Devasia

Table 23
Analysis based on sampling period.

Number of years	Number of articles	Reference number
< 1 Year	2	Zengin et al. (2011), Márquez-Vera et al. (2013a)
1 Year	13	Chuan et al. (2011), Dejaeger et al. (2012), Şen et al. (2012), Pratiwi (2013), Márquez-Vera et al. (2013b), Guarín et al. (2015), Sisovic et al. (2015), Lehr et al. (2016), Meedech et al. (2016), Costa et al. (2017), Kularbphettong (2017), Altaf et al. (2019), Karthikeyan et al. (2020)
1 and $\frac{1}{2}$ Year	2	Chen et al. (2014), Castro-Wunsch et al. (2017)
2 Years	4	Mashiloane and Mchunu (2013), Sanchez-Santillan et al. (2016), Al Breiki et al. (2019), Martins et al. (2019)
2 and $\frac{1}{2}$ Years	1	Jishan et al. (2015)
3 Years	9	Chau and Phung (2013), Palazuelos et al. (2013), Natek and Zwilling (2014), Devasia et al. (2016), Ayub et al. (2017), Rustia et al. (2018), Amazona and Hernandez (2019), Crivei et al. (2019), Mengash (2020)
4 Years	7	Tasnim et al. (2019), Vila et al. (2018), Ibrahim et al. (2018), Lottering et al. (2020), Utari et al. (2020), Santoso (2019), Jung (2018)
5 Years	7	Salinas and Stephens (2015), Leppänen et al. (2017), Jung (2016), Sukhija et al. (2018), Rojanavasu (2019), Tarmizi et al. (2019), Mkwazu and Yan (2020)
6 Years	3	Manhães et al. (2014), Pérez et al. (2018), M.Á. et al. (2020)
7 Years	1	Daud et al. (2017)
9 Years	2	Hoe et al. (2013), Chanlekha and Niramitranon (2018)
12 Years	3	Miguéis et al. (2018), Adekitan and Salau (2019), Ribeiro and Canedo (2020)

Table 24
Analysis based on size of dataset.

Size of Dataset	Number of Articles	Reference Number
<100	10	Chellatamilan et al. (2011), Cocea and Weibelzahl (2010), Wang and Liao (2011), Pathan et al. (2014), Bakaric et al. (2015), Shukor et al. (2015), Ramanathan et al. (2016), Pise and Kulkarni (2017), Abyaa et al. (2018), Rawat and Malhan (2019)
100 - 150	9	Kan et al. (2010), Natek and Zwilling (2014), Sorour et al. (2015), Martínez-Abad et al. (2018), Burgos et al. (2018), Rojanavasu (2019), Al Breiki et al. (2019), Akram et al. (2019), Ashraf et al. (2020)
151–200	12	Bodea et al. (2010), Göker et al. (2013), Palazuelos et al. (2013), Dangi and Srivastava (2014), Sisovic et al. (2015), Jishan et al. (2015), Kaur et al. (2015), Mayilvaganan and Kalpanadevi (2015), Sanchez-Santillan et al. (2016), Hamsa et al. (2016), Kassak et al. (2016), Figueira (2017)
201–250	3	Badr et al. (2016), Kaunang and Rotikan (2018), Kasthuriarachchi and Liyanage (2018)
251–300	4	Ahadi et al. (2015), Leppänen et al. (2017), Amazona and Hernandez (2019), Dimić et al. (2019)
301–350	3	Pratiwi (2013), Ayub et al. (2017), Lagus et al. (2018)
351–400	6	Mashiloane and Mchunu (2013), Athani et al. (2017), Tasnim et al. (2019), Umer et al. (2019), Rahman and Mahmud (2020), Agrawal et al. (2020)
401–450	3	Pruthi and Bhatia (2015), Barbosa Manhães et al. (2015), Rustia et al. (2018)
451–500	6	Amornsinlaphachai (2016), Chitra and Agrawal (2019), Ketui et al. (2019), Kamal and Ahuja (2019), Almutairi et al. (2019), Ajibade et al. (2018)
601–700	6	Márquez-Vera et al. (2013b), Márquez-Vera et al. (2013a), Devasia et al. (2016), Chaudhury et al. (2016), Stahovich and Lin (2016), Injadat et al. (2020a)
701–800	2	Daud et al. (2017), Ibrahim et al. (2018)
901–1000	5	Lehr et al. (2016), Hassan and Al-Razgan (2016), Buenaño-Fernández et al. (2017), Maitra et al. (2018) , Altaf et al. (2019)
1001–1500	9	Abaya and Gerardo (2013), Anh et al. (2014), Guo et al. (2015), Castro-Wunsch et al. (2017), Costa et al. (2017), Lagman et al. (2019), Crivei et al. (2019), Zaffar et al. (2018), El Aissaoui et al. (2020)
1501–2000	4	Blagojević and Micić (2013), Guarín et al. (2015), Adekitan and Salau (2019), Santoso (2019)
2001–2500	5	Hoe et al. (2013), Miguéis et al. (2018), Adekitan and Salau (2020), Utari et al. (2020), Mengash (2020)
2501–3000	5	Salinas and Stephens (2015), Agaoglu (2016), Srivastava et al. (2018), Spatiotis et al. (2018), Jung (2018)
3001–3500	2	Sen and Ucar (2012), Jung (2016)
4001–5000	4	El-Halees (2011), Şen et al. (2012), Martins et al. (2019), Lottering et al. (2020)
5001–6000	4	Chau and Phung (2013), Ragab et al. (2014), Ahmed et al. (2016), Kularbphettong (2017)
8000–10000	2	Chuan et al. (2011), Dejaeger et al. (2012)
10001–20000	2	Vila et al. (2018), Ab Ghani et al. (2018)
21000–25000	2	Chen et al. (2014), M.Á. et al. (2020)
31000–36000	3	Trandafili et al. (2012), Islam and Mahmud (2020), Ribeiro and Canedo (2020)

et al. (2016), Ayub et al. (2017), Rustia et al. (2018), Amazona and Hernandez (2019), Crivei et al. (2019), Mengash (2020), Tasnim et al. (2019), Vila et al. (2018), Ibrahim et al. (2018), Lottering et al. (2020), Utari et al. (2020), Santoso (2019), Jung (2018), and Salinas and Stephens (2015), Leppänen et al. (2017), Jung (2016), Sukhija et al. (2018), Rojanavasu (2019), Tarmizi et al. (2019), Mkwazu and Yan (2020) respectively. The research articles Hoe et al. (2013), Chanlekha and Niramitranon (2018), and Miguéis et al. (2018), Adekitan and Salau (2019), Ribeiro and Canedo (2020) worked on the dataset of sampling period 9 and 12 years. The research articles Manhães et al. (2014), Pérez et al. (2018), M.Á. et al. (2020), and Daud et al. (2017) considered the sampling period of 6 and 7 years respectively for dataset employed to build the model in EDM.

4.11. Analysis on basis of size of dataset

This subsection describes the analysis of research articles on basis of dataset size with the range <100, 100–150, 151–200, 201–250, 251–300, 301–350, 351–400, 401–450, 451–500, 601–700, 701–800, 901–1000, 1001–1500, 1501–2000, 2001–2500, 2501–3000, 3001–3500, 4001–5000, 5001–6000, 8000–10000, 10001–20000, 21000–25000, and 31000–36000. Table 24 describe the analysis of research articles based on size of dataset.

The research articles Chau and Phung (2013), Ragab et al. (2014), Ahmed et al. (2016), Kularbphettong (2017), Chuan et al. (2011), Dejaeger et al. (2012), Vila et al. (2018), Ab Ghani et al. (2018), Chen et al. (2014), M.Á. et al. (2020), and Trandafili et al. (2012), Islam and Mahmud (2020), Ribeiro and Canedo (2020) used the dataset in the range 5001–6000, 8000–10000, 10001–20000, 21000–25000, and 31000–36000 respectively. Twelve research articles Bodea et al.

(2010), Göker et al. (2013), Palazuelos et al. (2013), Dangi and Srivastava (2014), Sisovic et al. (2015), Jishan et al. (2015), Kaur et al. (2015), Mayilvaganan and Kalpanadevi (2015), Sanchez-Santillan et al. (2016), Hamsa et al. (2016), Kassak et al. (2016), Figueira (2017) considered the dataset size in the range 151–200 while the dataset size less than 100 is employed in the research articles Chellatamilan et al. (2011), Cocea and Weibelzahl (2010), Wang and Liao (2011), Pathan et al. (2014), Bakaric et al. (2015), Shukor et al. (2015), Ramanathan et al. (2016), Pise and Kulkarni (2017), Abyaa et al. (2018), and Rawat and Malhan (2019). The dataset size in the range 201–250, 251–300, 301–350, 351–400, 401–450, 451–500, 601–700, 701–800, and 901–1000 is observed in the research articles Badr et al. (2016), Kaunang and Rotikan (2018), Kasthuriarachchi and Liyanage (2018), Ahadi et al. (2015), Leppänen et al. (2017), Amazona and Hernandez (2019), Dimić et al. (2019), Pratiwi (2013), Ayub et al. (2017), Lagus et al. (2018), Mashiloane and Mchunu (2013), Athani et al. (2017), Tasnim et al. (2019), Umer et al. (2019), Rahman and Mahmud (2020), Agrawal et al. (2020), Pruthi and Bhatia (2015), Barbosa Manhães et al. (2015), Rustia et al. (2018), Amornsinlaphachai (2016), Chitra and Agrawal (2019), Ketui et al. (2019), Kamal and Ahuja (2019), Almutairi et al. (2019), Ajibade et al. (2018), Márquez-Vera et al. (2013b,a), Devasia et al. (2016), Chaudhury et al. (2016), Stahovich and Lin (2016), Injadat et al. (2020a), Daud et al. (2017), Ibrahim et al. (2018), and Lehr et al. (2016), Hassan and Al-Razgan (2016), Buenaño-Fernández et al. (2017), Maitra et al. (2018), Altaf et al. (2019) respectively in EDM.

4.12. Analysis on basis of data mining tools

This subsection analyzes the research articles based on data mining tool – H2O, KEEL, KNIME, Orange, RapidMiner, SAS and SPSS. Fig. 18 shows the analysis of research articles based on Data Mining tools. It is noted from Fig. 18 that forty seven research articles Bodea et al. (2010), Trandafili et al. (2012), Chau and Phung (2013), Göker et al. (2013), Márquez-Vera et al. (2013b,a), Mashiloane and Mchunu (2013), Palazuelos et al. (2013), Pathan et al. (2014), Ragab et al. (2014), Natek and Zwilling (2014), Pruthi and Bhatia (2015), Jacob et al. (2015), Parmar et al. (2015), Sisovic et al. (2015), Barbosa Manhães et al. (2015), Auddy and Mukhopadhyay (2015), Shukor et al. (2015), Kaur et al. (2015), Devasia et al. (2016), Lehr et al. (2016), Chaudhury et al. (2016), Sanchez-Santillan et al. (2016), Meedech et al. (2016), Hassan and Al-Razgan (2016), Ahmed et al. (2016), Ayub et al. (2017), Athani et al. (2017), Castro-Wunsch et al. (2017), Daud et al. (2017), Pise and Kulkarni (2017), Kaunang and Rotikan (2018), Chitra and Agrawal (2019), Abyaa et al. (2018), Martínez-Abad et al. (2018), Kularbphettong (2017), Al Breiki et al. (2019), Lagman et al. (2019), Chango et al. (2019), Al Fanah and Ansari (2019), Tarmizi et al. (2019), Rawat and Malhan (2019), Vila et al. (2018), Utari et al. (2020), Karthikeyan et al. (2020), and El Aissaoui et al. (2020) used the most popular data mining tool Weka for analyzing students' performance while RapidMiner tool is employed by eight research articles El-Halees (2011), Jishan et al. (2015), Angra and Ahuja (2017), Rustia et al. (2018), Rojanavasu (2019), Amazona and Hernandez (2019), Utari et al. (2020), and Santoso (2019). Data Mining tool H2O - a Java-based software for data modeling and general computing, KEEL, Orange, SAS/EnterpriseMiner, SPSS, and Watson Analytics is considered by the research articles Ribeiro and Canedo (2020), Chellatamilan et al. (2011), Adekitan and Salau (2020), Kamal and Ahuja (2019), Nasiri et al. (2012), and Pérez et al. (2018) respectively. KNIME tool is applied in the research articles Ibrahim et al. (2018), Adekitan and Salau (2019).

5. Comparative study of models for different EDM models

This section deals with comparison of different EDM datasets for various models in EDM based on performance metrics. This section covers the comparative study of models for predicting students' dropout rate, the comparative study of models for predicting students' performance in programming course, the comparative study of models for predicting students' placement and the comparative study of models for predicting students' performance in online education/ courses (see Table 25).

5.1. Comparative study of models for predicting students' dropout rate

This subsection considers the comparative study of models to predict students' dropout rate. Table 26 discusses the comparative study based on classification technique/s considered to build the model, best classification technique if model is build after comparing various classification algorithms based on various performance metrics, and the classification performance metrics such as Accuracy, True Positive (TP) rate, True Negative (TN) rate, Precision, Recall, Area under Curve (AUC), Receiver Operating Characteristic (ROC), F-measures, and cross validation K-fold. From Table 26, it is observed that 99% accuracy is obtained in research article (Lottering et al., 2020) in which four classification algorithms such as Decision Trees (e5.0), Support Vector Machine, Naive Bayes and Nearest Neighbour are compared on performance metrics Accuracy, Precision, Recall, and F-measures and found Support Vector Machine algorithm to be best to build the model for predicting the students' dropout rate. In research article (Tarmizi et al., 2019; Vila et al., 2018), the accuracy is 98% while 97% accuracy is observed in research articles Márquez-Vera et al. (2013b), and Burgos et al. (2018).

5.2. Comparative study of models for predicting students' performance in online/ e-learning/ distance education/ courses

This subsection considers the comparative study of models for predicting students' performance in online education/ courses. Table 27 describes the comparative study based on data mining/ classification technique/s considered to build the model, Best classification technique found if any, and the classification performance metrics such as Accuracy, True positive rate, False positive rate, Precision, F-measures, Correctly Classified Instances, K-fold, Error, Mean Absolute Error, t-Test, and Standard Deviation. From Table 27, it is found that 99% accuracy is observed for the classification algorithm Random Forests in the research article Figueira (2017).

The symbol '✓'in the table indicate that the parameter/s is/are used for deciding the performance of algorithm but the value of the performance metric is not provided in the research article.

5.3. Comparative study of models for predicting students' performance in programming course

This subsection considers the comparative study of models for predicting students' performance in programming course. Table 28 discusses the comparative study based on data mining/ classification technique/s considered to build the model, best classification technique if model is build after comparing various classification algorithms based on various performance metrics, and the classification performance metrics such as Accuracy, True Positive (TP) rate, True Negative (TN) rate, Precision, Recall, F-measures, cross validation K-fold and Mean Absolute Error. From Table 28, it is observed that 93% accuracy is found in the research article Márquez-Vera et al. (2013a) and 87% accuracy in the research article Pathan et al. (2014). The symbol '✓'in the table indicate that the parameter/s is/are used for deciding the performance of algorithm but the value of the performance metric is not provided in the research article.

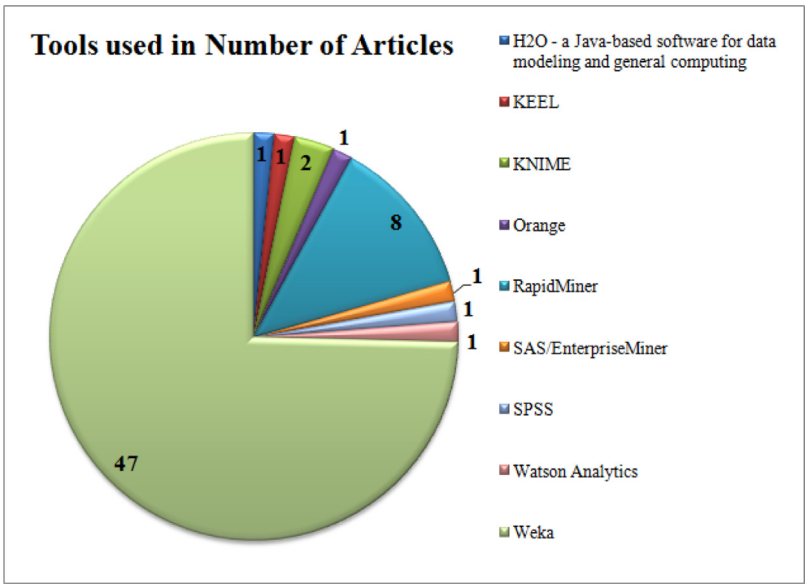


Fig. 18. Analysis on basis of tools used in EDM.

Table 25
EDM Models.

Sr. No.	Purpose of model	Ref.
1	Predicting the dropout rate	Márquez-Vera et al. (2013b), Palazuelos et al. (2013), Manhães et al. (2014), Barbosa Manhães et al. (2015), Meedech et al. (2016), Niu et al. (2018), Pérez et al. (2018), Burgos et al. (2018), Tasnim et al. (2019), Chango et al. (2019), Tarmizi et al. (2019), Vila et al. (2018) , Lottering et al. (2020) , M.Á. et al. (2020) , Santoso (2019), Ribeiro and Canedo (2020)
2	Predicting the students' performance in online/ e-learning/ Distance education/ courses	Bodea et al. (2010), Cocea and Weibelzahl (2010), Wang and Liao (2011), Nasiri et al. (2012), Sen and Ucar (2012), Salinas and Stephens (2015), Shukor et al. (2015), Kassak et al. (2016), Ayub et al. (2017), Buenaño-Fernández et al. (2017), Figueira (2017), Costa et al. (2017), Abyaa et al. (2018), Kularbphettong (2017), Al Fanah and Ansari (2019), El Aissaoui et al. (2020)
3	Analyzing students' performance in programming courses	Márquez-Vera et al. (2013a), Pathan et al. (2014), Ahadi et al. (2015), Sisovic et al. (2015), Amornsinlaphachai (2016), Badr et al. (2016), Leppänen et al. (2017), Costa et al. (2017), Lagus et al. (2018)
4	Analyzing the Learning Management System (LMS) data	Chellatamilan et al. (2011), Umer et al. (2019), Altaf et al. (2019), Dimić et al. (2019), Islam and Mahmud (2020), Ajibade et al. (2018), Injadat et al. (2020b)
5	Predicting the students' performance/ academic performance	Guruler et al. (2010), Sakurai et al. (2012), Bunkar et al. (2012), Bresfelean et al. (2012), Hoe et al. (2013), Chau and Phung (2013), Göker et al. (2013), Mashiloane and Mchunu (2013), Anh et al. (2014), Guo et al. (2015), Auddy and Mukhopadhyay (2015)
6	Predicting the students' placement	Şen et al. (2012), Pratiwi (2013), Gera and Goel (2015), Pruthi and Bhatia (2015), Ramanathan et al. (2016)

5.4. Comparative study of models for analyzing the Learning Management System (LMS) data

This subsection discusses about the comparative study of models for analyzing the Learning Management System (LMS) data. Table 29 discusses the comparative study of this model along with data mining/ classifications techniques used and performance parameter considered to check the performance of algorithm. From Table 29, it is noted that 93% accuracy is observed in the research article Ajibade et al. (2018). The symbol '✓'in the table indicate that the parameter/s is/are used for deciding the performance of algorithm but the value of the performance metric is not provided in the research article.

5.5. Comparative study of models for predicting the students' performance/ academic performance

This subsection illustrates the comparative study of models for predicting the students' performance/ academic performance. Table 30 describe the comparative study of this model for predicting the students' performance/ academic performance along with data mining/ classifications techniques used and performance parameters such as Accuracy, True positive (TP) rate, False positive (FP) rate, Precision, Recall, Receiver Operating Characteristic (ROC), F-measures, Correctly Classified Instances, Incorrectly Classified Instances, Cross-validation K-fold, Lift Value, and Cosine Cofficient considered to check the performance of algorithm. From Table 30, it is noted that 85% F-measures is observed in the research article Bunkar et al. (2012), Göker et al. (2013) while 94% accuracy is found in research article Chau and Phung (2013). The symbol '✓'in the table indicate that the parameter/s is/are used for deciding the performance of algorithm but the value of the performance metric is not provided in the research article.

5.6. Comparative study of models for predicting students' placement

This subsection considers the comparative study of models for predicting students' placement. Table 31 elucidates the comparative study based on classification technique/s considered to build the model, and the classification performance metrics such as Receiver Operating Characteristic (ROC), F-measures, Correctly Classified Instances, Incorrectly Classified Instances, Cross-validation K-fold, Relative absolute error, Root relative squared error, and Sum of Difference. From Table 26, it is noted that 95% accuracy is obtained in the research article Şen et al. (2012).

Table 26
Comparative study of the models to predict students' dropout rate.

Ref.	Classification Technique considered	Best Classification Technique found if any	Accuracy	TP rate	TN rate	Precision	Recall	AUC	ROC	F-measures	K-fold
Márquez-Vera et al. (2013b)	Jrip, Nnge, OneR, Prism, Ridor, J48, C4.5, SimpleCart, ADTree, Random Tree	ADTree	97%	98%	88%	–	–	–	–	–	10
Palazuelos et al. (2013)	J48, Random forests, Naíve Bayes, Bayesian networks, JRip, and Ridor	J48	70%	–	–	–	–	–	–	–	10
Manhães et al. (2014)	Naïve Bayes (NB), Multilayer Perceptron (MLP),Support Vector Machine with polynomial kernel (SVM1) and RBF kernel (SVM2) and Decision Table (DT)	Naíve Bayes	90%	76%	93%	–	–	–	–	–	10
Barbosa Manhães et al. (2015)	BayesNet (BN), J48 ,JRip (JR), Support Vector Machines, Multilayered Perceptrons (MLP), AdaBoost (AB), SimpleLogistic (SL), DecisionTable (DT), OneR , RandomForest (RF).	Multilayered Perceptrons (MLP), Simple Logistic (SL), Random Forest (RF), SVM	–	78%	–	82%	78%	–	–	–	–
Meedech et al. (2016)	JRip, OneR, Ridor, J48, SimpleCart, ADTree, RandomTree and REPTree	ADTree	80%	–	–	–	–	–	–	–	10
Niu et al. (2018)	Generalized Linear Models, Support Vector Machines, Tree Based Ensemble Learning Models, Neural Network Models, Naïve Bayes Models, K-Neighbours, Gaussian Progress. (Best-XGBoost)	XGBoost	94%	–	–	–	–	94%	–	–	4
Pérez et al. (2018)	Decision Trees, Logistic Regression, Naive Bayes and Random Forest	Random Forest	–	–	–	–	–	91%	–	–	5
Burgos et al. (2018)	Logistic regression	–	97%	–	–	99%	97%		–	–	10
Tasnim et al. (2019)	Logistic regression, Naive Bayes Classifier, Support Vector Machine and Threshold based approach for classification	–	–	–	–	96%	0.96	–	–	96%	5
Chango et al. (2019)	J48, REPTREE, RamdomTree, JRIP, Nnge, PART	PART	–	–	–	–	–	–	91%	81%	10
Tarmizi et al. (2019)	J48, Random Forest, Naïve Bayes (NB), SVM(RBF Kernal), SVM(Polynomial Kernal)	SVM(Polynomial Kernal)	98%	–	–	–	–	–	–	–	–
Vila et al. (2018)	RandomTree, Naïve Bayes	Random Tree	98%	98%	–	97%	98%	–	84%	97%	10
Lottering et al. (2020)	Decision Trees (e5.0), Support Vector Machine , Naive Bayes and Nearest Neighbour	Support Vector Machine	99%	–	–	98%	100%	–	–	99%	–
M.Á. et al. (2020)	Support Vector Machine	–	85%	–	–	–	–	–	–	–	–
Santoso (2019)	Decision tree (C4.5), and Naïve Bayes	Naíve Bayes	–	–	–	–	–	90%	–	–	10
Ribeiro and Canedo (2020)	GLM: Generalized Linear Model, GBM: Gradient Boosting Machine, SVM: Support Vector Machines, RF: Random Forest	Gradient Boosting Machine	–	–	–	–	–	–	86%	–	–

6. Current challenges, research gap and future direction

After reviewing 142 research articles from 2010–2020 publication year related to the EDM, some challenges/ limitations, research gap as well as future direction for researcher working in EDM are discussed in this section

6.1. Challenges

Following challenges are faced by the authors in the research related to the EDM

- Small dataset is available (Kan et al., 2010; Pathan et al., 2014; Guo et al., 2015; Badr et al., 2016; Kitanaka et al., 2017; Umer

Table 27
Comparative study of the models to predict students' placement.

Ref.	Data mining/ Classification techniques used	Best classification technique found if any	Accuracy	True positive rate	False positive rate	Precision	F-measures	Correctly classified instances	K-fold	Error	Mean absolute error	t-Test	Standard deviation
Bodea et al. (2010)	Simple K-Mean algorithm, J48 and PART	–	–	82%	6%	67%	73%	–	–	–	10%	–	–
Cocea and Weibelzahl (2010)	Bayesian Nets, Logistic regression (LR), Simple logistic classification, Instance-based classification with IBk algorithm, Attribute Selected Classification using J48 classifier and Best First search, Bagging using REP tree classifier, Classification via Regression, Decision Trees with J48 classifier	Simple logistic classification	87%	82%	9%	91%	–	–	–	21%	–	–	–
Wang and Liao (2011)	Artificial Neural Network	–	–	–	–	–	–	–	–	–	–	–	✓
Nasiri et al. (2012)	Regression and C5.0	–	89%	–	–	–	–	–	–	–	–	–	–
Sen and Ucar (2012)	Artificial neural networks, and Decision tree	–	98%	–	–	–	–	–	10	–	–	–	–
Salinas and Stephens (2015)	Naive Bayes	–	–	79%	–	–	–	–	–	–	–	–	–
Shukor et al. (2015)	C4.5	–	–	–	–	–	–	–	–	–	–	✓	-
Kassak et al. (2016)	Classifier based on polynomial regression and stochastic gradient descent	–	80%	–	–	–	–	–	–	–	–	-	-
Ayub et al. (2017)	Classification and Association Rule Mining	–	79%	–	–	–	–	–	10	–	–	–	–
Buenaño-Fernández et al. (2017)	Support Vector Machine (SVM).	–	75%	–	–	–	–	–	–	–	–	–	–
Figueira (2017)	Random Forests	–	99%	–	–	–	–	–	–	–	–	–	–
Costa et al. (2017)	Neural Networks, Decision Tree, Support Vector Machine, Naive Bayes	Support Vector Machine	–	–	–	–	92%	–	–	–	–	–	–
Abyaa et al. (2018)	Support Vector Machines (SVM), k-Nearest Neighbours (kNN), Naïve Bayes, Random forest, J48, Logistic regression, Bagging	–	–	–	–	–	–	✓	10	–	✓	–	–
Kularbphettong (2017)	J48, Bayesian networks	Bayesian networks	91%	–	–	91%	94%	-	–	–	-	–	–
Al Fanah and Ansari (2019)	Random Forests, Logistic Regressions and Bayesian Networks	Bayesian Networks	80%	–	–	–	–	–	10	–	–	–	–
El Aissaoui et al. (2020)	Naive Bayes, Cart, ID3, and C4.5	ID3	–	–	–	–	–	93%	10	–	–	–	–

et al., 2019; Altaf et al., 2019; Ab Ghani et al., 2018; Islam and Mahmud, 2020).

- Considering the data from various resources is required (El-Halees, 2011; Bunkar et al., 2012; Natek and Zwilling, 2014; Pruthi and Bhatia, 2015; Parmar et al., 2015; Bakaric et al., 2015; Kaur et al., 2015; Jung, 2016; Kaunang and Rotikan, 2018; Niu et al., 2018; Sukhija et al., 2018; Pérez et al., 2018; Burgos et al., 2018; Miguéis et al., 2018; Maitra et al., 2018; Al Breiki et al., 2019; Amazona and Hernandez, 2019; Martins et al., 2019; Adekitan and Salau, 2019; Utari et al., 2020).
- In addition to academic performance, some other factors such as social, economic, etc. needs to be considered (Márquez-Vera et al., 2013a; Shukor et al., 2015; Hassan and Al-Razgan, 2016; Buenaño-Fernández et al., 2017; Spatiotis et al., 2018; Martínez-Abad et al., 2018; Dimić et al., 2019; Vila et al., 2018; Kasthuriarachchi and Liyanage, 2018; Injadat et al., 2020a; Ajibade et al., 2018)
- Further optimization of model after applying data mining algorithms needs to be done (Kan et al., 2010; Blagojević and Micić, 2013; Guo et al., 2015; Rustia et al., 2018)
- Scope of dataset is limited to some range/ single source considered for collected data (Guruler et al., 2010; Guarín et al., 2015; Shukor et al., 2015; Miguéis et al., 2018; Tasnim et al.,

Table 28
Comparative study of the models to predict students' performance in programming course.

Ref.	Data mining/ Classification technique used	Best classification technique found if any	Accuracy	True positive rate	True Negatives rate	Precision	Recall	F-measures	Cross-validation K-fold	Mean Absolute Error
Márquez-Vera et al. (2013a)	Jrip, Nnge, OneR, Prism, Ridor, J48, C4.5, SimpleCart, ADTree, Random Tree, REPTree, Interpretable Classification Rule Mining(ICRM)	Interpretable Classification Rule Mining	93%	93%	88%	–	–	–	10	–
Pathan et al. (2014)	ID3 and C4.5	–	87%	–	–	–	–	–	–	–
Ahadi et al. (2015)	Naive Bayes, Bayesian Network, Decision Table, Conjuctive Rule, PART, ADTree, J48, Random Forest, Decision Stum	Random Forest	75%	–	–	–	–	–	10	–
Sisovic et al. (2015)	C4.5, JRip and PART	C4.5	83%						10	
Amornsin-laphachai (2016)	Artificial Neural Network, K-Nearest Neighbour, Naive Bayes, Bayesian Belief Network, JRIP, ID3 and C4.5	C4.5	–	–	–	74%	75%	74%	10	24%
Badr et al. (2016)	Classification and Association Rule Mining	–	67%	–	–	–	–	–	–	–
Leppänen et al. (2017)	support vector methods: a linear kernel, an RBF kernel and a sigmoid kernel, Regression analysis	–	✓	–	–	–	–	–	10	–
Costa et al. (2017)	Neural Networks, Decision Tree, Support Vector Machine, Naive Bayes	Support Vector Machine	–	–	–	–	–	92%	–	–
Lagus et al. (2018)	Support Vector Machine, Random Forest, and AdaBoost	Random Forest	–	–	–	✓	✓	✓	10	–

Table 29
Comparative Study of Models for analyzing the Learning Management System (LMS) data.

New Ref.	Data mining/ Classification technique used	Accu-racy	True posi-tive (TP) rate	False posi-tive (FP) rate	Preci-sion	Recall	Receiver Operating Character-istic (ROC)	F-measures	Cross-validation K-fold	Root mean square error (RMSE)	Relative absolute error (RAE)	Root relative squared error	Standard Deviation
Chel-latamilan et al. (2011)	J48 and K-means	–	100%	11%	88%	100%	96%	94%	3	23%	21%	46%	–
Umer et al. (2019)	Random Forest classifier (RF), Naïve Bayes (NB), Logistic regression (LR), Linear Discriminating analysis (LDA)	–	–	–	–	–	–	–	10	–	–	–	19.7
Altaf et al. (2019)	Multi-Layer Perceptions	84%	–	–	–	83%	–	–	–	–	–	–	–
Dimić et al. (2019)	Naïve Bayes, Hidden Naíve Bayes, J48 and Random Forest classifiers	–	94%	1%	91%	–	–	–	–	–	–	–	–
Islam and Mahmud (2020)	Logistic Regression (LR), k-nearest neighbours (KNN) and Decision Trees	63%	–	–	–	–	–	–	–	–	–	–	–
Ajibade et al. (2018)	Naïve Bayesian (NB), Decision Tree (DT), K-Nearest Neighbour (KNN), Discriminant Analysis (Disc) and Pairwise Coupling (PWC). Ensemble techniques – AdaBoost, Bag and RUSBoost	94%	–	–	90%	91%	–	91%	10	–	–	–	–
Injadat et al. (2020b)	K-nearest neighbour (k-NN), random forest (RF), Support Vector machine (SVM), Logistic Regression (LR), Multi-Layer Perceptron (MLP), and Naïve Bayes (NB).	✓	–	–	✓	–	–	✓	3				

2019; Akram et al., 2019; Vila et al., 2018; Lottering et al., 2020; Mengash, 2020)

- Appropriate data/ biased data will be collected if students/ learners learned that it is used for evaluation (El-Halees, 2011)
- Human efforts for data analysis and interpretation is required (Chen et al., 2014)
- Very few attributes in the dataset is the major concerned (Bodea et al., 2010; Bunkar et al., 2012)

Table 30
Comparative Study of Models for Predicting the students' performance/ academic performance.

New Ref.	Data Mining/ Classification Technique used	Accuracy	True positive (TP) rate	False positive (FP) rate	Precision	Recall	Receiver Operating-Characteristic (ROC)	F-measures	Correctly Classified Instances	Incorrectly Classified Instances	Cross-validation K-fold	Lift Value	Cosine Cofficient
Guruler et al. (2010)	Decision Tree	–	–	–	–	-	–	–	–	–	-	✓	–
Sakurai et al. (2012)	Decision Tree	–	–	–	–	–	–	–	–	–	–	–	✓
Bunkar et al. (2012)	ID3 (Iterative Dichotomiser 3), C4.5 and CART	–	86%	11%	85%	86%	92%	85%	–	–	–	–	-
Bresfelean et al. (2012)	k-means clustering and C4.5, REPTree	–	–	–	–	–	–	-	79%	-	10		–
Hoe et al. (2013)	Classification using CHAID (Chi-Squared Automatic Interaction Detection)	70%	–	–	–	–	–	–	-	-	-	-	–
Chau and Phung (2013)	Naïve Bayes, Support Vector Machine (SVM), Neural Network, Logistic Regression, Knearest Neighbour (k-nn), and Decision Tree C4.5, Random Forest	94%	–	–	–	–	99%	–	-	-	10	-	–
Göker et al. (2013)	Naive Bayes, J48, Bays Net, RBF Network	85%	–	–	85%	85%	87%	85%	85%	15%	–	-	–
Mashiloane and Mchunu (2013)	J48 classifier, Naíve Bayes and Decision Table	88%	–	27%	89%	–	–	-	86%	14%	–	-	–
Anh et al. (2014)	Naïve Bayes, Neural Network, Support Vector Machine, K-Nearest Neighbour (K-nn), Decision Tree C4.5, and Random Forest.	✓	–	–	–	–	✓	–	-	–	5	-	–
Guo et al. (2015)	Naïve Bayes, Multilayer Perception (MLP) and SVM ,SPPN Prediction Network	–	–	–	77%	–	–	-	-	-	-	-	–
Auddy and Mukhopadhyay (2015)	J48, JRip, QuickRules Fuzzy-Rough rule induction, Fuzzy Nearest Neighbour, Fuzzy Rough Nearest Neighbour and Vaguely Quantified Nearest Neighbour (VQNN)	–	78%	25%	78%	78%	78%	77%	–	–	10	–	–

Table 31
Comparative study of the models to predict students' placement.

Ref.	Classification techniques considered	Accuracy	Receiver Operating Characteristic (ROC)	F-measures	Correctly Classified Instances	Incorrectly Classified Instances	Cross-validation K-fold	Relative absolute error	Root relative squared error	Sum of Difference
Şen et al. (2012)	C5 Decision Tree	95%	–	–	–	–	10	–	–	–
Pratiwi (2013)	J48, SimpleCart, Kstar, SMO, Naive Bayes, OneR	80%			80%	20%	10	–	–	–
Gera and Goel (2015)	Decision Tree	–	85%	25%	–	–	–	80%	89%	–
Pruthi and Bhatia (2015)	J48 Decision Tree algorithm,NaiVe Bayes	62%	–	–	–	–	–	-	–	
Ramanathan et al. (2016)	Decision Tree and Random Forest	–	–	–	–	–	–	–	–	>4

6.2. Limitations

Following are the limitations found after reviewing the research articles related to the EDM

- Data processing techniques are required to prepare the dataset to build the model in EDM (Devasia et al., 2016; Jung, 2016; Kiu, 2018; Tasnim et al., 2019; Akram et al., 2019; Amazona and Hernandez, 2019; M.Á. et al., 2020).
- Few data mining techniques were applied to build the model so there is need to apply more data mining techniques for the developed/ proposed system. (Kan et al., 2010; Nasiri et al., 2012; Bresfelean et al., 2012; Abaya and Gerardo, 2013; Chau

and Phung, 2013; Márquez-Vera et al., 2013b; Mashiloane and Mchunu, 2013; Palazuelos et al., 2013; Dangi and Srivastava, 2014; Pathan et al., 2014; Chaudhury et al., 2016; Sanchez-Santillan et al., 2016; Meedech et al., 2016; Athani et al., 2017; Kiu, 2018; Patil et al., 2018; Kaunang and Rotikan, 2018; Lagus et al., 2018; Rustia et al., 2018; Martínez-Abad et al., 2018; Sukhija et al., 2018; Miguéis et al., 2018; Maitra et al., 2018; Tasnim et al., 2019; Amazona and Hernandez, 2019; Altaf et al., 2019; Dimić et al., 2019; Vila et al., 2018; Ab Ghani et al., 2018; Yousafzai et al., 2020; Utari et al., 2020; Rahman and Mahmud, 2020)

- Combining various data mining techniques for analyzing the students' performance is required (Chellatamilan et al., 2011; Alshehri, 2019; Dimić et al., 2019)
- Few parameters were used in determining the performance of algorithm (Guruler et al., 2010; Zengin et al., 2011; Sakurai et al., 2012; Dejaeger et al., 2012)
- Only one parameter is used to determine the performance of the designed model (Guruler et al., 2010; Chuan et al., 2011; Wang and Liao, 2011; Sakurai et al., 2012; Hoe et al., 2013; Pathan et al., 2014; Manhães et al., 2014; Natek and Zwilling, 2014; Parmar et al., 2015; Guo et al., 2015; Shukor et al., 2015; Badr et al., 2016; Hassan and Al-Razgan, 2016; Castro-Wunsch et al., 2017; Buenaño-Fernández et al., 2017; Figueira, 2017; Jung, 2016; Costa et al., 2017; Srivastava et al., 2018; Patil et al., 2018; Chitra and Agrawal, 2019; Lagman et al., 2019; Adekitan and Salau, 2020; Tarmizi et al., 2019; Zaffar et al., 2018; Adekitan and Salau, 2019; Yahya, 2019; M.Á. et al., 2020; Mengash, 2020; Rahman and Mahmud, 2020; Mkwazu and Yan, 2020; Islam and Mahmud, 2020; Agrawal et al., 2020)
- The work mentioned in research articles Shukor et al. (2015), Lagus et al. (2018), Tasnim et al. (2019), Lottering et al. (2020), Mengash (2020) is limited to applying the method to the dataset which cannot be generalized.

6.3. Research gap

The research gap in EDM noted from 142 research articles are –

- Combination of data mining techniques that are least used in research are -
  - Classification and Association Rule Mining (3.5% of total 142 research articles);
  - Classification, Clustering and Association Rule Mining (2.11% of total 142 research articles);
  - Classification, Regression, and Clustering (1.4% of total 142 research articles); and
  - Classification and Ensemble technique (0.7% of total 142 research articles).
- Least used classifier as per Weka are Meta(5.63% of total 142 research articles), Rules (14.08% of total 142 research articles) and Lazy (18.3% of total 142 research articles).
- Least used classification algorithms in the range of 0.7%–5% of total 142 research articles, are Simple Logistic, Linear Discriminating analysis, Artificial Neural Network, Sequential minimal optimization (SMO), Linear Regression, Multiple Regression, RBF Network, IBk (Instance based Learner), Kstar, QuickRules Fuzzy-Rough ruleinduction, Fuzzy Nearest Neighbour, Fuzzy Rough Nearest Neighbour, Vaguely Quantified Nearest Neighbour (VQNN), AdaBoost, Attribute Selected Classifier, Bagging, Classification Vis Regression, Logit Boot, Decision Table, Conjuctive Rules, OneR, PART, Prism, RIDOR, Nnge, Decision Stump, M5P, ADTree, Random Tree, REPTree, CART (Classification and Regression Trees), SimpleCart, C5.0, Oblique Classifier, XGBoost, Gradient boosted tree, Gaussian Processes Random Tree (GPRT), Classification using CHAID (Chi-Squared Automatic Interaction Detection), Discriminant analysis, GLM: Generalized Linear Model, Singular Value Decomposition (SVD), Interpretable Classification Rule Mining, MaxMargin Multi-Label (M3L) classifier, LIBSVM, Cross-out classifier, and Classifier based on polynomial regression and stochastic gradient descent.
- Least used clustering techniques are Expectation Maximization and K-modes.
- Least used association rule algorithms are FP-growth and Relational association rule algorithm.
- Least used performance parameters for classification algorithms in the range of 0.7%–1.4% of total research article 142, are Notch difference, Lift Value, False Negatives, t-Test, and Cosine Coefficient.
- Fifty six (39.43% of total 142 research articles) research articles considered the dataset size less than 500 while only six (4.22% of total 142 research articles) research articles Trandafili et al. (2012), Chen et al. (2014), Vila et al. (2018), Ab Ghani et al. (2018), M.Á. et al. (2020), Islam and Mahmud (2020), and Ribeiro and Canedo (2020) used the dataset size more than 10,000.

6.4. Improvements required in the proposed/ developed models in reviewed research articles

- More instances and/ attributes needs to be added in dataset (Bunkar et al., 2012; Palazuelos et al., 2013; Pruthi and Bhatia, 2015; Devasia et al., 2016; Ramanathan et al., 2016; Chitra and Agrawal, 2019; Francis and Babu, 2019; Crivei et al., 2019; Kasthuriarachchi and Liyanage, 2018; Islam and Mahmud, 2020)
- Analysis of data from different educational system and application of other data mining techniques needs to be considered (Nasiri et al., 2012; Trandafili et al., 2012)
- Application different data mining techniques are required (Bresfelean et al., 2012; Mashiloane and Mchunu, 2013; Pathan et al., 2014; Sanchez-Santillan et al., 2016; Jung, 2016; Costa et al., 2017; Chitra and Agrawal, 2019; Kularbphettong, 2017; Ketui et al., 2019; Chango et al., 2019; Kamal and Ahuja, 2019; Vila et al., 2018)
- Other classification techniques are needed to enhance the performance of model (Bunkar et al., 2012; Abaya and Gerardo, 2013; Palazuelos et al., 2013; Dangi and Srivastava, 2014; Natek and Zwilling, 2014; Athani et al., 2017; Kaunang and Rotikan, 2018; Pérez et al., 2018; Rojanavasu, 2019)
- Other clustering techniques are required (Ramanathan et al., 2016; Mengash, 2020)
- Capability of resulting classifier requires to be improved (Chau and Phung, 2013)
- Combination of data mining algorithms can be applied (Lagman et al., 2019)
- More data/ Data from large group of students/ students from other areas/ other institutions will be required (Trandafili et al., 2012; Palazuelos et al., 2013; Pathan et al., 2014; Chen et al., 2014; Natek and Zwilling, 2014; Parmar et al., 2015; Ahadi et al., 2015; Barbosa Manhães et al., 2015; Amornsinlaphachai, 2015; Meedech et al., 2016; Badr et al., 2016; Castro-Wunsch et al., 2017; Buenaño-Fernández et al., 2017; Kitanaka et al., 2017; Costa et al., 2017; Rustia et al., 2018; Abyaa et al., 2018; Pérez et al., 2018; Rojanavasu, 2019; Lagman et al., 2019; Amazona and Hernandez, 2019; Chango et al., 2019; Altaf et al., 2019; Crivei et al., 2019; Yousafzai et al., 2020; Adekitan and Salau, 2019; Mengash, 2020; Injadat et al., 2020a)
- More training samples should be gathered (Guo et al., 2015)
- Data processing techniques should be included (Devasia et al., 2016)
- Removal of imbalance data is required to enhance the model (Tasnim et al., 2019; Dimić et al., 2019)

- Feature selection method should be considered to use the best features of dataset (Chango et al., 2019; Rahman and Mahmud, 2020)
- Deep learning based classifier can be used for performance prediction of students (Yousafzai et al., 2020)
- Ensemble methods for classification of data sets can be used (Rawat and Malhan, 2019)

6.5. Future direction

Inspite of the research gap, future direction for working in EDM are mentioned below-

- Combination of data mining techniques that are mostly used in research are classification and clustering algorithm.
- Mostly used classifier as per Weka are Bays (47.88% of total 142 research articles), Function (33.1% of total 142 research articles), Trees (61.26% of total 142 research articles), and PMML classifier (32.39% of total 142 research articles).
- Mostly used classification algorithms are Naïve Bayes(44.36% of total 142 research articles), Support Vector Machine (SVM) (28.16% of total 142 research articles), Random Forest (26.76% of total 142 research articles), J48 (16.9% of total 142 research articles), C4.5(15.49% of total 142 research articles), K-Nearest Neighbour (KNN) (14.7% of total 142 research articles), Logistic (14.7% of total 142 research articles), Neural Network (10.56% of total 142 research articles), and Multilayer Perceptron (MLP) (11.26% of total 142 research articles).
- Mostly used clustering techniques is K-means clustering algorithm.
- Mostly used association rule algorithm is Apriori association rule algorithm.
- Naive Bays, Support vector machine, Random forest and Decision Tree are found to be the best classification techniques after comparing various classification algorithms using performance parameters such as accuracy, precision, recall, F-measures, k-fold value, etc.
- Performance parameters of Classification algorithms such as accuracy (58.45% of total 142 research articles), precision (30.98% of total 142 research articles), recall (26.76% of total 142 research articles), F-measures (26.05% of total 142 research articles), and k-fold value (46.48% of total 142 research articles) are found in most of research articles.

7. Conclusion

In this review articles, analysis of 142 research articles from publication year 2010–2020 downloaded from various research databases such as IEEE, Elsevier, Springer and ACM, is discussed. Also current research in last two years (2021 and 2022) is also discussed. The use of classification techniques and classification techniques along with other techniques in EDM are presented thoroughly. For 142 research articles reviewed, the analysis is done based on number of articles related to various data mining techniques from various research database, Publication Year of research articles, research database, Number of Citations, and Number Pages in Research Article. Also the comparative study is considered for Classification Techniques along with its' combination with other data mining techniques. Analysis in terms of Yearwise Number of Research Articles employing Classification Techniquein EDM; Classification with other Data Mining Technique used in EDM; classifier as per Weka Tool; Classification Techniques; Clustering Techniques; Association Rule Techniques; Selecting the best Classification Technique; Classification performance metric; software used in EDM; Sampling Period; dataset size; and data mining tools are illustrated. Analysis of research articles based on performance parameters such as Analysis on basis of Accuracy, TP rate, FP rate, Precision, Recall, F-measures, Cross Validation, Correctly Classified Instance, Incorrectly Classified Instance, RMSE (Root Mean Square Error), MAE (Mean absolute error), AUC (Area under the ROC Curve), and ROC (Receiver operating characteristic curve) are also mentioned in this review article.

Classification techniques are found to be mostly used technique for classifying and analyzing students' performance in EDM. It is found that Naïve Bays, Random Forest, Support vector machine and J48 are mostly considered classification techniques. The classification algorithms Naïve Bays, Random Forest and Support Vector Machine are noted to be the best classification algorithms after comparing various classification algorithms based on various performance parameters such as accuracy, precision, recall, etc. The parameters accuracy, precision, recall, f-measures and k-fold value are used by most of research articles. Data mining tools considered for analysis of students' performance are Weka, and RapidMiner while software used to build the model in EDM are Java, R and Python programming languages. Classification algorithms under the classifiers as per Weka tool such as Tree, Bays, Function and PMML (Predictive model markup language) classifier are applied in most of the research articles.

Research gaps in the articles are identified in terms of least used classification algorithms, dataset size considered, sampling period, etc. Future direction for researcher working in EDM to predict students' performance are discussed so that more useful research from students' point of view in EDM can be carried out to improve the teaching–learning process in education sector.

CRediT authorship contribution statement

Sunita M. Dol: Writing – original draft. **Pradip M. Jawandhiya:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Ab Ghani, N.L., Cob, Z.C., Drus, S.M., Sulaiman, H., 2018. Student enrolment prediction model in higher education institution: A data mining approach. In: International Symposium of Information and Internet Technology. Springer, Cham, pp. 43–52.

Abaya, S.A., Gerardo, B.D., 2013. An education data mining tool for marketing based on C4. 5 classification technique. In: 2013 Second International Conference on E-Learning and E-Technologies in Education. ICEEE, IEEE, pp. 289–293.

Abualigah, L., Diabat, A., 2022. Chaotic binary group search optimizer for feature selection. Expert Syst. Appl. 192, 116368.

Abualigah, L.M., Khader, A.T., 2017. Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. J. Supercomput. 1–23.

Abualigah, L.M., Khader, A.T., Al-Betar, M.A., 2016. Unsupervised feature selection technique based on genetic algorithm for improving the Text Clustering. In: Computer Science and Information Technology (CSIT), 2016 7th International Conference on. IEEE, pp. 1–6.

Abualigah, L.M., Khader, A.T., Hanandeh, E.S., 2018. A hybrid strategy for krill herd algorithm with harmony search algorithm to improve the data clustering. Intell. Dec. Technol. 12 (1), 3–14.

Abyaa, A., Idrissi, M.K., Bennani, S., 2018. Predicting the learner's personality from educational data using supervised learning. In: Proceedings of the 12th International Conference on Intelligent Systems: Theories and Applications. pp. 1–7.

Adekitan, A.I., Salau, O., 2019. The impact of engineering students' performance in the first three years on their graduation result using educational data mining. Heliyon 5 (2), e01250.

Adekitan, A.I., Salau, O., 2020. Toward an improved learning process: the relevance of ethnicity to data mining prediction of students' performance. SN Appl. Sci. 2 (1), 1–15.

Agaoglu, M., 2016. Predicting instructor performance using data mining techniques in higher education. IEEE Access 4, 2379–2387.

Agrawal, R., Singh, J., Ghosh, S.M., 2020. Performance appraisal of an educational institute using data mining techniques. In: Computing in Engineering and Technology. Springer, Singapore, pp. 733–745.

Ahadi, A., Lister, R., Haapala, H., Vihavainen, A., 2015. Exploring machine learning methods to automatically identify students in need of assistance. In: Proceedings of the Eleventh Annual International Conference on International Computing Education Research. pp. 121–130.

Ahmed, A.M., Rizaner, A., Ulusoy, A.H., 2016. Using data mining to predict instructor performance. Procedia Comput. Sci. 102, 137–142.

Ajibade, S.S.M., Ahmad, N.B., Shamsuddin, S.M., 2018. A data mining approach to predict academic performance of students using ensemble techniques. In: International Conference on Intelligent Systems Design and Applications. Springer, Cham, pp. 749–760.

Akram, A., Fu, C., Li, Y., Javed, M.Y., Lin, R., Jiang, Y., Tang, Y., 2019. Predicting students' academic procrastination in blended learning course using homework submission data. IEEE Access 7, 102487-102498.

Al Breiki, B., Zaki, N., Mohamed, E.A., 2019. Using educational data mining techniques to predict student performance. In: 2019 International Conference on Electrical and Computing Technologies and Applications. ICECTA, IEEE, pp. 1–5.

Al Fanah, M., Ansari, M.A., 2019. Understanding e-learners' behaviour using data mining techniques. In: Proceedings of the 2019 International Conference on Big Data and Education. pp. 59–65.

Aldosari, F., Abualigah, L., Almotairi, K.H., 2022. A normal distributed dwarf mongoose optimization algorithm for global optimization and data clustering applications. Symmetry 14 (5), 1021.

Almutairi, S., Shaiba, H., Bezbradica, M., 2019. Predicting students' academic performance and main behavioral features using data mining techniques. In: International Conference on Computing. Springer, Cham, pp. 245–259.

Alomari, O.A., Khader, A.T., Al-Betar, M.A., Abualigah, L.M., 2017. MRMR BA: a hybrid gene selection algorithm for cancer classification. J. Theor. Appl. Inf. Technol. 95 (12), 2610–2618.

Alshehri, Y.A., 2019. Applying explanatory analysis in education using different regression methods. In: Proceedings of the 2019 4th International Conference on Information and Education Innovations. pp. 109–115.

AlShourbaji, I., Kachare, P., Zogaan, W., Muhammad, L.J., Abualigah, L., 2022. Learning features using an optimized artificial neural network for breast cancer diagnosis. SN Comput. Sci. 3 (3), 1–8.

Altaf, S., Soomro, W., Rawi, M.I.M., 2019. Student performance prediction using multilayers artificial neural networks: A case study on educational data mining. In: Proceedings of the 2019 3rd International Conference on Information System and Data Mining. pp. 59–64.

Amazona, M.V., Hernandez, A.A., 2019. Modelling student performance using data mining techniques: Inputs for academic program development. In: Proceedings of the 2019 5th International Conference on Computing and Data Engineering. pp. 36–40.

Amornsinlaphachai, P., 2015. The design of a framework for cooperative learning through web utilizing data mining technique to group learners. Proc.-Soc. Behav. Sci. 174, 27–33.

Amornsinlaphachai, P., 2016. Efficiency of data mining models to predict academic performance and a cooperative learning model. In: 2016 8th International Conference on Knowledge and Smart Technology. KST, IEEE, pp. 66–71.

Angra, S., Ahuja, S., 2017. Implementation of data mining algorithms on student's data using rapid miner. In: 2017 International Conference on Big Data Analytics and Computational Intelligence. ICBDAC, IEEE, pp. 387–391.

Anh, N.T.M., Chau, V.T.N., Phung, N.H., 2014. Towards a robust incomplete data handling approach to effective educational data classification in an academic credit system. In: 2014 International Conference on Data Mining and Intelligent Computing. ICDMIC, IEEE, pp. 1–7.

Ashraf, M., Zaman, M., Ahmed, M., 2020. An intelligent prediction system for educational data mining based on ensemble and filtering approaches. Procedia Comput. Sci. 167, 1471–1483.

Athani, S.S., Kodli, S.A., Banavasi, M.N., Hiremath, P.S., 2017. Student academic performance and social behavior predictor using data mining techniques. In: 2017 International Conference on Computing, Communication and Automation. ICCCA, IEEE, pp. 170–174.

Attiya, I., Abualigah, L., Elsadek, D., Chelloug, S.A., Abd Elaziz, M., 2022. An intelligent chimp optimizer for scheduling of IoT application tasks in fog computing. Mathematics 10 (7), 1100.

Auddy, A., Mukhopadhyay, S., 2015. Data mining on ICT usage in an academic campus: a case study. In: International Conference on Distributed Computing and Internet Technology. Springer, Cham, pp. 443–447.

Ayub, M., Toba, H., Wijanto, M.C., Yong, S., 2017. Modelling online assessment in management subjects through educational data mining. In: 2017 International Conference on Data and Software Engineering (ICoDSE). IEEE, pp. 1–6.

Badr, G., Algobail, A., Almutairi, H., Almutery, M., 2016. Predicting students' performance in university courses: a case study and tool in KSU mathematics department. Procedia Comput. Sci. 82, 80–89.

Bakaric, M.B., Matetic, M., Sisovic, S., 2015. Text mining student reports. In: Proceedings of the 16th International Conference on Computer Systems and Technologies. pp. 382–389.

Barbosa Manhães, L.M., da Cruz, S.M.S., Zimbrão, G., 2015. Towards automatic prediction of student performance in STEM undergraduate degree programs. In: Proceedings of the 30th Annual ACM Symposium on Applied Computing. pp. 247–253.

Blagojević, M., Micić, Ž., 2013. A web-based intelligent report e-learning system using data mining techniques. Comput. Electr. Eng. 39 (2), 465–474.

Bodea, C.N., Bodea, V., Mogos, R., 2010. Student performance in online project management courses: A data mining approach. In: World Summit on Knowledge Society. Springer, Berlin, Heidelberg, pp. 470–479.

Bresfelean, V.P., Bresfelean, M., Lacurezeanu, R., 2012. Data mining tasks in a student-oriented dss. In: Advanced Information Technology in Education. Springer, Berlin, Heidelberg, pp. 321–328.

Buenaño-Fernández, D., Luján-Mora, S., Villegas-Ch, W., 2017. Improvement of massive open online courses by text mining of students' emails: a case study. In: Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality. pp. 1–7.

Bunkar, K., Singh, U.K., Pandya, B., Bunkar, R., 2012. Data mining: Prediction for performance improvement of graduate students using classification. In: 2012 Ninth International Conference on Wireless and Optical Communications Networks. WOCN, IEEE, pp. 1–5.

Burgos, C., Campanario, M.L., de la Peña, D., Lara, J.A., Lizcano, D., Martínez, M.A., 2018. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Comput. Electr. Eng. 66, 541–556.

Castro-Wunsch, K., Ahadi, A., Petersen, A., 2017. Evaluating neural networks as a method for identifying students in need of assistance. In: Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education. pp. 111–116.

Chango, W., Cerezo, R., Romero, C., 2019. Predicting academic performance of university students from multi-sources data in blended learning. In: Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems. pp. 1–5.

Chanlekha, H., Niramitranon, J., 2018. Student performance prediction model for early-identification of at-risk students in traditional classroom settings. In: Proceedings of the 10th International Conference on Management of Digital EcoSystems. pp. 239–245.

Chau, V.T.N., Phung, N.H., 2013. Imbalanced educational data classification: An effective approach with resampling and random forest. In: The 2013 RIVF International Conference on Computing & Communication Technologies-Research, Innovation, and Vision for Future. RIVF, IEEE, pp. 135–140.

Chaudhury, P., Mishra, S., Tripathy, H.K., Kishore, B., 2016. Enhancing the capabilities of student result prediction system. In: Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. pp. 1–6.

Chellatamilan, T., Ravichandran, M., Suresh, R.M., Kulanthaivel, G., 2011. Effect of mining educational data to improve adaptation of learning in e-learning system.

Chen, X., Vorvoreanu, M., Madhavan, K., 2014. Mining social media data for understanding students' learning experiences. IEEE Trans. Learn. Technol. 7 (3), 246–259.

Chitra, Jalota, Agrawal, Rashmi, 2019. Analysis of educational data mining using classification. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). IEEE.

Chuan, H., Ruifan, L., Yixin, Z., 2011. Combining different classifiers in educational data mining. In: International Conference on Applied Informatics and Communication. Springer, Berlin, Heidelberg, pp. 467–473.

Cocea, M., Weibelzahl, S., 2010. Disengagement detection in online learning: Validation studies and perspectives. IEEE Trans. Learn. Technol. 4 (2), 114–124.

Costa, E.B., Fonseca, B., Santana, M.A., de Araújo, F.F., Rego, J., 2017. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Comput. Hum. Behav. 73, 247–256.

Crivei, L.M., Czibula, G., Mihai, A., 2019. A study on applying relational association rule mining based classification for predicting the academic performance of students. In: International Conference on Knowledge Science, Engineering and Management. Springer, Cham, pp. 287–300.

Dabhade, P., Agarwal, R., Alameen, K.P., Fathima, A.T., Sridharan, R., Gopakumar, G., 2021. Educational data mining for predicting students' academic performance using machine learning algorithms. Mater. Today: Proc. 47, 5260–5267.

Dangi, A., Srivastava, S., 2014. Educational data classification using selective Naïve Bayes for quota categorization. In: 2014 IEEE International Conference on MOOC, Innovation and Technology in Education. MITE, IEEE, pp. 118–121.

Daud, A., Aljohani, N.R., Abbasi, R.A., Lytras, M.D., Abbas, F., Alowibdi, J.S., 2017. Predicting student performance using advanced learning analytics. In: Proceedings of the 26th international conference on world wide web companion. pp. 415–421.

Dejaeger, K., Goethals, F., Giangreco, A., Mola, L., Baesens, B., 2012. Gaining insight into student satisfaction using comprehensible data mining techniques. European J. Oper. Res. 218 (2), 548–562.

Devasia, T., Vinushree, T.P., Hegde, V., 2016. Prediction of students performance using educational data mining. In: 2016 International Conference on Data Mining and Advanced Computing. SAPIENCE, IEEE, pp. 91–95.

Dimić, G., Rančić, D., Pronić-Rančić, O., Milošević, D., 2019. An approach to educational data mining model accuracy improvement using histogram discretization and combining classifiers into an ensemble. In: Smart Education and E-Learning 2019. Springer, Singapore, pp. 267–280.

El Aissaoui, O., El Madani, Y.E.A., Oughdir, L., Dakkak, A., El Allioui, Y., 2020. Mining learners' behaviors: An approach based on educational data mining techniques. In: Embedded Systems and Artificial Intelligence. Springer, Singapore, pp. 655–670.

El-Halees, A., 2011. Mining opinions in user-generated contents to improve course evaluation. In: International Conference on Software Engineering and Computer Systems. Springer, Berlin, Heidelberg, pp. 107–115.

Figueira, Á., 2017. Mining Moodle logs for grade prediction: a methodology walk-through. In: Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality. pp. 1–8.

Francis, B.K., Babu, S.S., 2019. Predicting academic performance of students using a hybrid data mining approach. J. Med. Syst. 43 (6), 1–15.

Gera, M., Goel, S., 2015. A model for predicting the eligibility for placement of students using data mining technique. In: International Conference on Computing, Communication & Automation. IEEE, pp. 114–117.

Göker, H., Bülbül, H.I., Irmak, E., 2013. The estimation of students' academic success by data mining methods. In: 2013 12th International Conference on Machine Learning and Applications. vol. 2, IEEE, pp. 535–539.

Guarín, C.E.L., Guzmán, E.L., González, F.A., 2015. A model to predict low academic performance at a specific enrollment using data mining. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje 10 (3), 119–125.

Guo, B., Zhang, R., Xu, G., Shi, C., Yang, L., 2015. Predicting students performance in educational data mining. In: 2015 International Symposium on Educational Technology. ISET, IEEE, pp. 125–128.

Guruler, H., Istanbullu, A., Karahasan, M., 2010. A new student performance analysing system using knowledge discovery in higher educational databases. Comput. Educ. 55 (1), 247–254.

Hamsa, H., Indiradevi, S., Kizhakkethottam, J.J., 2016. Student academic performance prediction model using decision tree and fuzzy genetic algorithm. Proc. Technol. 25, 326–332.

Hassan, S.M., Al-Razgan, M.S., 2016. Pre-university exams effect on students GPA: a case study in IT department. Procedia Comput. Sci. 82, 127–131.

Hernández-Leal, E., Duque-Méndez, N.D., Cechinel, C., 2021. Unveiling educational patterns at a regional level in Colombia: data from elementary and public high school institutions. Heliyon 7 (9), e08017.

Hoe, A.C.K., Ahmad, M.S., Hooi, T.C., Shanmugam, M., Gunasekaran, S.S., Cob, Z.C., Ramasamy, A., 2013. Analyzing students records to identify patterns of students' performance. In: 2013 International Conference on Research and Innovation in Information Systems. ICRIIS, IEEE, pp. 544–547.

Hussain, S., Ayoub, M., Jilani, G., Yu, Y., Khan, A., Wahid, J.A., et al., 2022. Aspect2Labels: A novelistic decision support system for higher educational institutions by using multi-layer topic modelling approach. Expert Syst. Appl. 209, 118119.

Ibrahim, Z.M., Bader-El-Den, M., Cocea, M., 2018. Mining unit feedback to explore students' learning experiences. In: UK Workshop on Computational Intelligence. Springer, Cham, pp. 339–350.

Injadat, M., Moubayed, A., Nassif, A.B., Shami, A., 2020a. Multi-split optimized bagging ensemble model selection for multi-class educational data mining. Appl. Intell. 50 (12), 4506–4528.

Injadat, M., Moubayed, A., Nassif, A.B., Shami, A., 2020b. Systematic ensemble model selection approach for educational data mining. Knowl.-Based Syst. 200, 105992.

Islam, S., Mahmud, H., 2020. Integration of learning analytics into learner management system using machine learning. In: Proceedings of the 2020 2nd International Conference on Modern Educational Technology. pp. 1–4.

Jacob, John, et al., 2015. Educational data mining techniques and their applications. In: 2015 International Conference on Green Computing and Internet of Things (ICGCIoT). IEEE.

Jishan, S.T., Rashu, R.I., Mahmood, A., Billah, F., Rahman, R.M., 2015. Application of optimum binning technique in data mining approaches to predict students' final grade in a course. In: Computational Intelligence in Information Systems. Springer, Cham, pp. 159–170.

Jung, E., 2016. A comparison of data mining methods in analyzing educational data. In: Advances in Computer Science and Ubiquitous Computing. Springer, Singapore, pp. 173–178.

Jung, E., 2018. An educational data mining with Bayesian networks for analyzing variables affecting parental attachment. In: Advances in Computer Science and Ubiquitous Computing. Springer, Singapore, pp. 557–563.

Kamal, P., Ahuja, S., 2019. Academic performance prediction using data mining techniques: Identification of influential factors effecting the academic performance in undergrad professional course. In: Harmony Search and Nature Inspired Optimization Algorithms. Springer, Singapore, pp. 835–843.

Kan, L., Xingyuan, X., Ping, L., 2010. DMCMS: A data mining based course management system. In: 2010 Second International Workshop on Education Technology and Computer Science. vol. 3, IEEE, pp. 145–148.

Kapgate, D., 2022. Efficient quadcopter flight control using hybrid SSVEP+ P300 visual brain computer interface. Int. J. Hum.–Comput. Int. 38 (1), 42–52.

Karthikeyan, V.G., Thangaraj, P., Karthik, S., 2020. Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation. Soft Comput. 24 (24), 18477–18487.

Kassak, O., Kompan, M., Bielikova, M., 2016. Student behavior in a web-based educational system: Exit intent prediction. Eng. Appl. Artif. Intell. 51, 136–149.

Kasthuriarachchi, K.T.S., Liyanage, S.R., 2018. Predicting students' academic performance using utility based educational data mining. In: International Conference on Frontier Computing. Springer, Singapore, pp. 29–39.

Kaunang, F.J., Rotikan, R., 2018. Students' academic performance prediction using data mining. In: 2018 Third International Conference on Informatics and Computing. ICIC, IEEE, pp. 1–5.

Kaur, P., Singh, M., Josan, G.S., 2015. Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia Comput. Sci. 57, 500–508.

Ketui, N., Wisomka, W., Homjun, K., 2019. Using classification data mining techniques for students performance prediction. In: 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON). IEEE, pp. 359–363.

Kitanaka, Y., Takeuchi, K., Hirokawa, S., 2017. Predicting learning result of learner in e-learning course with feature selection using SVM. In: Proceedings of the 2017 9th International Conference on Education Technology and Computers. pp. 122–125.

Kiu, C.C., 2018. Data mining analysis on student's academic performance through exploration of student's background and social activities. In: 2018 Fourth International Conference on Advances in Computing, Communication & Automation. ICACCA, IEEE, pp. 1–5.

Kularbphettong, K., 2017. Analysis of students' behavior based on educational data mining. In: Proceedings of the Computational Methods in Systems and Software. Springer, Cham, pp. 167–172.

Lagman, A.C., Calleja, J.Q., Fernando, C.G., Gonzales, J.G., Legaspi, J.B., Ortega, J.H.J.C., et al., 2019. Embedding Naïve Bayes algorithm data model in predicting student graduation. In: Proceedings of the 3rd International Conference on Telecommunications and Communication Engineering. pp. 51–56.

Lagus, J., Longi, K., Klami, A., Hellas, A., 2018. Transfer-learning methods in programming course outcome prediction. ACM Trans. Comput. Edu. 18 (4), 1–18.

Lehr, S., Liu, H., Kinglesmith, S., Konyha, A., Robaszewska, N., Medinilla, J., 2016. Use educational data mining to predict undergraduate retention. In: 2016 IEEE 16th International Conference on Advanced Learning Technologies. ICALT, vol. 42, IEEE, pp. 428–430.

Leppänen, L., Leinonen, J., Ihantola, P., Hellas, A., 2017. Predicting academic success based on learning material usage. In: Proceedings of the 18th Annual Conference on Information Technology Education. pp. 13–18.

Lottering, R., Hans, R., Lall, M., 2020. A model for the identification of students at risk of dropout at a university of technology. In: 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (IcABCD). IEEE, pp. 1–8.

M.Á., Prada, Domínguez, M., Vicario, J.L., Alves, P.A.V., Barbu, M., Podpora, M., et al., 2020. Educational data mining for tutoring support in higher education: A web-based tool case study in engineering degrees. IEEE Access 8, 212818-212836.

Ma, X., Qu, J.H., Xu, H.M., Ling, Y.T., 2021. E-learning performance prediction based on attention mechanism. In: 2021 the 6th International Conference on Distance Education and Learning. pp. 152–156.

Maitra, S., Madan, S., Kandwal, R., Mahajan, P., 2018. Mining authentic student feedback for faculty using Naïve Bayes classifier. Procedia Comput. Sci. 132, 1171–1183.

Malini, J., Kalpana, Y., 2021. Investigation of factors affecting student performance evaluation using education materials data mining technique. Mater. Today: Proc. 47, 6105–6110.

Manhães, L.M.B., da Cruz, S.M.S., Zimbrão, G., 2014. WAVE: an architecture for predicting dropout in undergraduate courses using EDM. In: Proceedings of the 29th Annual Acm Symposium on Applied Computing. pp. 243–247.

Márquez-Vera, C., Cano, A., Romero, C., Ventura, S., 2013a. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Appl. Intell. 38 (3), 315–330.

Márquez-Vera, C., Morales, C.R., Soto, S.V., 2013b. Predicting school failure and dropout by using data mining techniques. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje 8 (1), 7–14.

Martínez-Abad, F., Gamazo, A., Rodríguez-Conde, M.J., 2018. Big data in education: detection of ICT factors associated with school effectiveness with data mining techniques. In: Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality. pp. 145–150.

Martins, M.P., Miguéis, V.L., Fonseca, D.S.B., Alves, A., 2019. A data mining approach for predicting academic success–a case study. In: International Conference on Information Technology & Systems. Springer, Cham, pp. 45–56.

Mashiloane, L., Mchunu, M., 2013. Mining for marks: a comparison of classification algorithms when predicting academic performance to identify "students at risk". In: Mining Intelligence and Knowledge Exploration. Springer, Cham, pp. 541–552.

Mayilvaganan, M., Kalpanadevi, D., 2015. Cognitive skill analysis for students through problem solving based on data mining techniques. Procedia Comput. Sci. 47, 62–75.

Meedech, P., Iam-On, N., Boongoen, T., 2016. Prediction of student dropout using personal profile and data mining approach. In: Intelligent and Evolutionary Systems. Springer, Cham, pp. 143–155.

Mengash, H.A., 2020. Using data mining techniques to predict student performance to support decision making in university admission systems. IEEE Access 8, 55462–55470.
Miguéis, V.L., Freitas, A., Garcia, P.J., Silva, A., 2018. Early segmentation of students according to their academic performance: A predictive modelling approach. Decis. Support Syst. 115, 36–51.
Mkwazu, H.R., Yan, C., 2020. Grade prediction method for university course selection based on decision tree. In: Proceedings of the 2020 International Conference on Aviation Safety and Information Technology. pp. 593–599.
Mostafa, R.R., Ewees, A.A., Ghoniem, R.M., Abualigah, L., Hashim, F.A., 2022. Boosting chameleon swarm algorithm with consumption AEO operator for global optimization and feature selection. Knowl.-Based Syst. 246, 108743.
Nasiri, M., Minaei, B., Vafaei, F., 2012. Predicting GPA and academic dismissal in LMS using educational data mining: A case mining. In: 6th National and 3rd International Conference of e-Learning and e-Teaching. IEEE, pp. 53–58.
Natek, S., Zwilling, M., 2014. Student data mining solution–knowledge management system related to higher education institutions. Expert Syst. Appl. 41 (14), 6400–6407.
Niu, Z., Li, W., Yan, X., Wu, N., 2018. Exploring causes for the dropout on massive open online courses. In: Proceedings of ACM Turing Celebration Conference-China. pp. 47–52.
Palazuelos, C., García-Saiz, D., Zorrilla, M., 2013. Social network analysis and data mining: An application to the e-learning context. In: International Conference on Computational Collective Intelligence. Springer, Berlin, Heidelberg, pp. 651–660.
Parmar, K., Vaghela, D., Sharma, P., 2015. Performance prediction of students using distributed data mining. In: 2015 International Conference on Innovations in Information, Embedded and Communication Systems. ICIIECS, IEEE, pp. 1–5.
Pathan, A.A., Hasan, M., Ahmed, M.F., Farid, D.M., 2014. Educational data mining: A mining model for developing students' programming skills. In: The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014). IEEE, pp. 1–5.
Patil, R., Salunke, S., Kalbhor, M., Lomte, R., 2018. Prediction system for student performance using data mining classification. In: 2018 Fourth International Conference on Computing Communication Control and Automation. ICCUBEA, IEEE, pp. 1–4.
Pérez, B., Castellanos, C., Correal, D., 2018. Predicting student drop-out rates using data mining techniques: A case study. In: IEEE Colombian Conference on Applications in Computational Intelligence. Springer, Cham, pp. 111–125.
Pise, N., Kulkarni, P., 2017. Evolving learners' behavior in data mining. Evol. Syst. 8 (4), 243–259.
Pratiwi, O.N., 2013. Predicting student placement class using data mining. In: Proceedings of 2013 IEEE International Conference on Teaching, Assessment and Learning for Engineering. TALE, IEEE, pp. 618–621.
Pruthi, K., Bhatia, P., 2015. Application of data mining in predicting placement of students. In: 2015 International Conference on Green Computing and Internet of Things (ICGCIoT). IEEE, pp. 528–533.
Ragab, A.H.M., Noaman, A.Y., Al-Ghamdi, A.S., Madbouly, A.I., 2014. A comparative analysis of classification algorithms for students college enrollment approval using data mining. In: Proceedings of the 2014 Workshop on Interaction Design in Educational Environments. pp. 106–113.
Rahman, M., Mahmud, A., 2020. Classification on educational performance evaluation dataset using feature extraction approach. In: Proceedings of the International Conference on Computing Advancements. pp. 1–6.
Ramanathan, L., Geetha, A., Khalid, M., Swarnalatha, P., 2016. Apply of sum of difference method to predict placement of students' using educational data mining. In: Information Systems Design and Intelligent Applications. Springer, New Delhi, pp. 367–377.
Rawat, K.S., Malhan, I.V., 2019. A hybrid classification method based on machine learning classifiers to predict performance in educational data mining. In: Proceedings of 2nd International Conference on Communication, Computing and Networking. Springer, Singapore, pp. 677–684.
Ribeiro, R.C., Canedo, E.D., 2020. Using data mining techniques to perform school dropout prediction: A case study. In: 17th International Conference on Information Technology–New Generations (ITNG 2020). Springer, Cham, pp. 211–217.
Rojanavasu, P., 2019. Educational data analytics using association rule mining and classification. In: 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON). IEEE, pp. 142–145.
Rustia, R.A., Cruz, M.M.A., Burac, M.A.P., Palaoag, T.D., 2018. Predicting student's board examination performance using classification algorithms. In: Proceedings of the 2018 7th International Conference on Software and Computer Applications. pp. 233–237.
Sakurai, Y., Takada, K., Tsuruta, S., Knauf, R., 2012. A case study on using data mining for university curricula. In: 2012 IEEE 12th International Conference on Advanced Learning Technologies. IEEE, pp. 3–4.
Salinas, J.G.M., Stephens, C.R., 2015. Applying data mining techniques to identify success factors in students enrolled in distance learning: a case study. In: Mexican International Conference on Artificial Intelligence. Springer, Cham, pp. 208–219.
Sanchez-Santillan, M., Paule-Ruiz, M., Cerezo, R., Nuñez, J., 2016. Predicting students' performance: Incremental interaction classifiers. In: Proceedings of the Third (2016) ACM Conference on Learning@ Scale. pp. 217–220.
Santoso, L.W., 2019. The analysis of student performance using data mining. In: Advances in Computer Communication and Computational Sciences. Springer, Singapore, pp. 559–573.
Sen, B., Ucar, E., 2012. Evaluating the achievements of computer engineering department of distance education students with data mining methods. Proc. Technol. 1, 262–267.
Şen, B., Uçar, E., Delen, D., 2012. Predicting and analyzing secondary education placement-test scores: A data mining approach. Expert Syst. Appl. 39 (10), 9468–9476.
Shukor, N.A., Tasir, Z., Van der Meijden, H., 2015. An examination of online learning effectiveness using data mining. Proc.-Soc. Behav. Sci. 172, 555–562.
Sisovic, S., Matetic, M., Bakaric, M.B., 2015. Mining student data to assess the impact of moodle activities and prior knowledge on programming course success. In: Proceedings of the 16th International Conference on Computer Systems and Technologies. pp. 366–373.
Sorour, S., Goda, K., Mine, T., 2015. Correlation of topic model and student grades using comment data mining. In: Proceedings of the 46th ACM Technical Symposium on Computer Science Education. pp. 441–446.
Spatiotis, N., Perikos, I., Mporas, I., Paraskevas, M., 2018. Evaluation of an educational training platform using text mining. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence. pp. 1–5.
Srivastava, S., Karigar, S., Khanna, R., Agarwal, R., 2018. Educational data mining: Classifier comparison for the course selection process. In: 2018 International Conference on Smart Computing and Electronic Enterprise. ICSCEE, IEEE, pp. 1–5.
Stahovich, T.F., Lin, H., 2016. Enabling data mining of handwritten coursework. Comput. Graph. 57, 31–45.
Sukhija, K., Aggarwal, N., Jindal, M., 2018. EDARC: Collaborative frequent pattern and analytical mining tool for exploration of educational information. In: Recent Findings in Intelligent Computing Techniques. Springer, Singapore, pp. 251–259.
Tarmizi, S.S.A., Mutalib, S., Hamid, N.H.A., Abdul-Rahman, S., Ab Malik, A.M., 2019. A case study on student attrition prediction in higher education using data mining techniques. In: International Conference on Soft Computing in Data Science. Springer, Singapore, pp. 181–192.
Tasnim, N., Paul, M.K., Sattar, A.S., 2019. Identification of drop out students using educational data mining. In: 2019 International Conference on Electrical, Computer and Communication Engineering. ECCE, IEEE, pp. 1–5.
Teoh, C.W., Ho, S.B., Dollmat, K.S., Chai, I., 2022. An evolutionary algorithm-based optimization ensemble learning model for predicting academic performance. In: 2022 11th International Conference on Software and Computer Applications. pp. 102–107.
Trandafili, E., Allkoçi, A., Kajo, E., Xhuvani, A., 2012. Discovery and evaluation of student's profiles with machine learning. In: Proceedings of the Fifth Balkan Conference in Informatics. pp. 174–179.
Umer, R., Mathrani, A., Susnjak, T., Lim, S., 2019. Mining activity log data to predict student's outcome in a course. In: Proceedings of the 2019 International Conference on Big Data and Education. pp. 52–58.
Utari, M., Warsito, B., Kusumaningrum, R., 2020. Implementation of data mining for drop-out prediction using random forest method. In: 2020 8th International Conference on Information and Communication Technology (ICoICT). IEEE, pp. 1–5.
Veluri, R.K., Patra, I., Naved, M., Prasad, V.V., Arcinas, M.M., Beram, S.M., Raghuvanshi, A., 2022. Learning analytics using deep learning techniques for efficiently managing educational institutes. Mater. Today: Proc. 51, 2317–2320.
Vila, D., Cisneros, S., Granda, P., Ortega, C., Posso-Yépez, M., García-Santillán, I., 2018. Detection of desertion patterns in university students using data mining techniques: A case study. In: International Conference on Technology Trends. Springer, Cham, pp. 420–429.
Vinker, E., Rubinstein, A., 2022. Mining code submissions to elucidate disengagement in a computer science MOOC. In: LAK22: 12th International Learning Analytics and Knowledge Conference. pp. 142–151.
Wang, S., Hussien, A.G., Jia, H., Abualigah, L., Zheng, R., 2022. Enhanced remora optimization algorithm for solving constrained engineering optimization problems. Mathematics 10 (10), 1696.
Wang, Y.H., Liao, H.C., 2011. Data mining for adaptive learning in a TESL-based e-learning system. Expert Syst. Appl. 38 (6), 6480–6485.
Xu, H., Qu, J., Ma, X., Ling, Y., 2021. Prediction and visualization of academic procrastination in online learning. In: 2021 the 6th International Conference on Distance Education and Learning. pp. 133–139.
Yahya, A.A., 2019. Swarm intelligence-based approach for educational data classification. J. King Saud Univ.-Comput. Inf. Sci. 31 (1), 35–51.
Yousafzai, B.K., Hayat, M., Afzal, S., 2020. Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. Educ. Inf. Technol. 25 (6), 4677–4697.
Zaffar, M., Hashmani, M.A., Savita, K.S., 2018. Comparing the performance of FCB, chi-square and relief-f filter feature selection algorithms in educational data mining. In: International Conference of Reliable Information and Communication Technology. Springer, Cham, pp. 151–160.
Zengin, K., Esgi, N., Erginer, E., Aksoy, M.E., 2011. A sample study on applying data mining research techniques in educational science: Developing a more meaning of data. Proc.-Soc. Behav. Sci. 15, 4028–4032.