International Journal of Advanced Technology and Engineering Exploration, Vol 8(84)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2021.874521

Review Article

A systematic literature review on student performance predictions

Hasnah Nawang[1*], Mokhairi Makhtar[2] and Wan Mohd Amir Fazamin Wan Hamzah[3]
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Terengganu, Malaysia[1]
School of Computer Science, Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Tembila Campus, Terengganu, Malaysia[2]
School of Information Technology, Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Tembila Campus, Terengganu, Malaysia[3]

Received: 31-July-2021; Revised: 01-November-2021; Accepted: 06-November-2021

Abstract

Prediction of student performance in educational institutions is a major topic of debate among researchers in efforts to improve teaching and learning. Effective prediction techniques and features would help educators and teachers design appropriate teaching content to help learners study according to predicted outcomes. The purpose of this paper is to present a systematic literature review on predictions of students' performance in higher education institutions and secondary schools using Machine Learning, Educational Data Mining, and Learning Analytics methodologies. The review used in this study was designed to: i) provide an overview of techniques and algorithms used to predict students' performance; and ii) identify the features that have the greatest impact on students' performance. This paper also outlined several future insights in terms of applying hybrid techniques to educational datasets in order to improve accuracy in predicting students' performance.

Keywords

Educational data mining, Machine learning, Learning analytics, Students, Performance prediction.

1.Introduction

Nowadays, many high school and higher educational systems generate a large number of student information through the learning management system, examination data, students' activities, library system, etc. [1]. This situation leads to increases in the volume and types of educational data in every institution. Machine learning, learning analytics, and data mining approaches have been widely used on educational data to predict students' performance. These approaches have shown that several techniques and algorithms are useful in understanding this domain, which is poorly accessed by human capability.

Students' success has become an important metric to higher educational institutes as well as secondary level schools. In higher education institutions, students' performance plays a vital role in determining their job success [2]. Good academic performance assures employers of a candidate's quality and reliability.

In order to build and apply a predictive model, features that correlate with the value to be predicted must be collected and processed. There are many features affecting students' performance, and they can be divided into a few groups, such as students' previous education, students' e-learning activity, demographics features [3], students' social network information [4], behaviour variables [5], external assessment, extra-curricular activities and academic performance [6], school design [7], parental involvement, etc. [8]. As there are numerous features and approaches are utilised to forecast students' performance, this study provides a thorough evaluation of student performance predictions for high/secondary schools and higher-level institutions in terms of the most influential attributes and methods regularly employed by researchers.

Either in university/college or secondary school, students' performance prediction had drawn much attention not only amongst educators but the machine learning community as well. Predicting students' performance has become a challenging task due to the increased amount of data contained in educational

*Author for correspondence

systems [9, 10]. The students' interest in learning are varies. Some may be able to learn independently using e-learning, while others may need to study with the assistance of educators who prepare their teaching materials based on the needs of the students, while others may perform adequately in their examinations with the full facilities provided in school or institutions, and those with a higher family income may be able to find outside sources to fund extra-classes for their children's educational needs. The variety of this interest has motivated the authors to identify what features mostly contribute to learning outcomes together with which suitable machine learning techniques that will later assist institutions administration in coming up with solutions in offering the best education environment to their students. Educators, on the other hand, will be able to construct instructional content based on students' acceptance and lead students' learning progress.

In the recent years, with the advances of the application of technologies to forecasting students' performance, there are still gaps to be filled in order to discover most commonly used features and techniques. Despite of other review papers on student performance exist, few of them emphasize on the importance of combining multiple groups of features and techniques in order to enhance the accuracy percentage. Previous literature review on students' performance [2, 11] has discussed this topic on general and did not emphasize on the impact of hybrid approaches that able to help in optimizing the accuracy. Thus, we aim to comprehensively map, analyze and review the articles that have been published in 2016 to 2020.

2.Research methodology

The conduct of this review paper had followed the recommended procedures on performing a systematic review as provided by Kitchenham et al. [12]. The suggested procedure for a review is divided into three stages, which are:

1. Planning the review
2. Conducting the review
3. Reporting the review

2.1Planning the review

This systematic review was conducted due to the need to summarise the latest five years of research on student performance prediction. As the first phase of the review is planning, it is thus necessary to identify the importance of this review. This systematic review was proposed in order to support the objectives of this study which are:

1. To summarise existing methodologies used in student performance prediction.
2. To identify the most common features used in predicting students' performance.
3. To identify the most common algorithms used to predict students' performance.
4. To identify the gaps in previous research

After the objectives of the research were identified, the most important activity in a systematic review protocol which is to formulate the research questions was then conducted. Based on Kitchenham, 2009 recommendations [12], the construction of research questions must consider three viewpoints of the study criteria, which are population, intervention, and outcomes. The following are details for each study criteria: 1) population: high school and higher educational institutes; 2) intervention: methods, algorithms, and techniques for prediction; and 3) outcome: best features or variables used as performance predictors and successful prediction techniques or approaches. Hence, the following study criteria have led to the mapping of this study research questions (RQ):

1. What are the features commonly used by researchers to predict students' performance?
2. Which methods that are commonly used to investigate students' performance?
3. What are the best algorithms or techniques used in student performance prediction?

2.2Conducting the review

2.2.1Dataset

To ensure the adaptation of relevant papers in this review, papers were collected from four databases namely IEEE Xplore, ScienceDirect, SpringerLink, and EBSCOhost. The four databases are as illustrated in *Table 1*. The search was conducted in September 2021.

Table 1 Databases used to search for papers

No.	Database name	No of papers
1	IEEEXplore	10
2	ScienceDirect	60
3	Springer Link	50
4	EbcoHost	26

2.2.2Search strategy

The search strategy was used to identify relevant papers to the review. The following phrases - Machine learning, educational data mining, learning analytics, students' performance analysis, predictive model, students' performance prediction, students' academic performance, school, high school, higher education - were used to ensure that the selected papers only discussed machine learning, educational data mining, and learning analytics approaches. The search terms have been constructed by identifying the techniques of information extraction, level of education, and category of prediction. We used Boolean operators like AND, OR, and NOT in our search strings.

2.2.3Selection criteria

In this review, the study selection criteria were planned in order to identify the primary studies based on the following parts: 1) studies that used machine learning, educational data mining or learning analytics technique(s) for predictive modelling; 2) studies in peer-reviewed journals or conference proceedings written in English; and 3) studies that investigated the prediction of students' performance at higher educational or high school levels. This systematic review was performed on studies published from the years 2016 to 2020.

2.2.4The inclusion and exclusion criteria

The main objective is to include as many articles as possible starting from 2016 to 2021 that are related to the research questions. Then, certain criteria were set to identify whether an article should be included or not in the analysis.

Inclusion criteria

1. Research papers on Students Performance Prediction
2. Papers from 2016 to 2020 era.
3. Papers written in English.
4. Not a review paper

Exclusion Criteria

1. Studies not using machine learning or data mining techniques
2. Duplicates paper
3. Irrelevant titles, abstract and keywords
4. Not in peer-reviewed

Figure 1 and illustrates the PRISMA flowchart used as a guide during the selection process. As discussed in Section 2.2.1, four databases, which are IEEE Xplore, EBSCOhost, ScienceDirect, and SpringerLink, were used to search for papers during this systematic review. Ten items were retrieved from IEEE Xplore, 26 items were from EBSCOhost, 60 items were from ScienceDirect, and 50 items were from SpringerLink, making a total of 146 records. Firstly, published papers that match the search strings were identified. Secondly, papers were selected based on the title, abstract, and keywords relevant to the eligible criteria. The final selection was done by reading the full content of the papers. This study selected 40 articles for the subsequent review process as shown in *Figure 1*. The first screening removed 40 records due to problems such as duplicate papers, papers not written in English, and papers that were not peer-reviewed, leaving a total of 106 records. The second screening removed 30 papers due to irrelevant titles, abstract, and keywords used, bringing the new total to 76 records. Finally, the full text paper reading process excluded another 36 papers, making a total number of 40 papers remaining for analysis. These papers were tabulated based on their Paper Id and author/s name in *Table 2*.

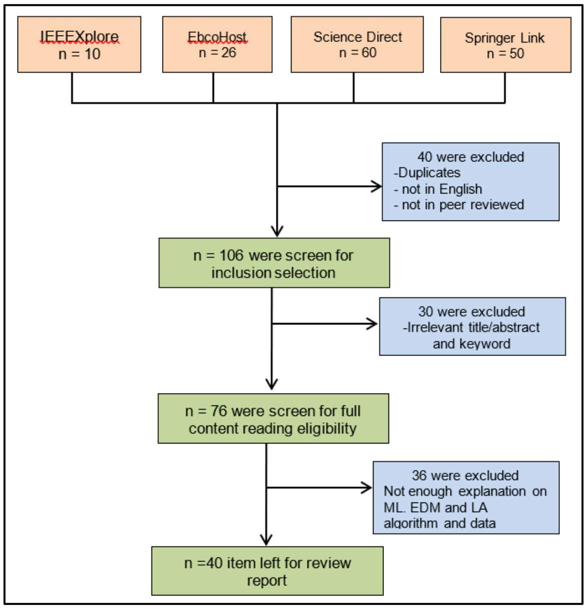
IEEEXplore
n = 10
EbcoHost
n = 26
Science Direct
n = 60
Springer Link
n = 50
40 were excluded
-Duplicates
- not in English
- not in peer reviewed
n = 106 were screen for inclusion selection
30 were excluded
-Irrelevant title/abstract and keyword
n = 76 were screen for full content reading eligibility
36 were excluded
Not enough explanation on ML. EDM and LA algorithm and data
n =40 item left for review report

Figure 1 PRISMA flowchart for research extraction and direction

Table 2 Selected papers for study review

Paper Id	Author	Paper Id	Author
1	[13]	P21	[33]
P2	[14]	P22	[34]
P3	[15]	P23	[35]
P4	[16]	P24	[36]
P5	[17]	P25	[37]
P6	[18]	P26	[38]
P7	[19]	P27	[39]
P8	[20]	P28	[40]
P9	[21]	P29	[41]
P10	[22]	P30	[42]
P11	[23]	P31	[43]
P12	[24]	P32	[44]
P13	[25]	P33	[45]
P14	[26]	P34	[46]
P15	[27]	P35	[47]
P16	[28]	P36	[48]

Paper Id	Author	Paper Id	Author
P17	[29]	P37	[49]
P18	[30]	P38	[50]
P19	[31]	P39	[51]
P20	[32]	P40	[52]

3.Results

This section will discuss the on the analysis of features, methods and algorithms used in this review study. There are two main factors that contribute to success in student performance prediction, which are the features or attributes of educational data and the techniques or algorithms used to explore educational data in order to make predictions and patterns [53].

3.1(RQ:i) What are the features commonly used by researchers to predict students' performance?

Different researchers have found different attributes that contribute to prediction accuracy in their studies. However, some attributes carry the same meaning but are differently labelled, and those attributes can be divided into a few groups, such as students' previous education, students' e-learning activities, demographic features, students' social network information, behaviour variables, extra-curricular activities, school design, academic performance, parental involvement, etc.

During the full text reading process, some features were identified and grouped for this review. Six groups of features were detected and consisted of Demographic Features, E-Learning Features, Social Network Features, School Design Features, Academic Performance Features, and Previous Education Features.

There are 40 primary studies included in this review. As shown in *Figure 2*, amongst the findings, academic performance group features had the highest percentage of recurrence in terms of researchers using them to predict students' performance with the 87.5% followed by previous education group with 47.5% and demographic features with 0.5% difference that is 47%. While the recurrence of percentage for social network features is 10% and the lowest is the school design features with 5% of recurrence. From the percentage, it shows that most researchers find academic performance features [11], such as GPA to be significant predictors of students' performance.

Figure 3 show the percentage of studies using one or more number *and Table 3* groups the features found into the six categories.
Figure 4 illustrated the list of attributes used in this study based on their feature groups.

Table 4 will go detailed on the best attributes discover by the researchers in this review.

Using more than one category of features seems popular in predicting students' performance as shown in *Figure 2*, 47% researchers in this review used the combination of two different groups of features in their study, followed by three combination of features groups with 23% and 10% of the study used four combination of features groups. As for only one category of feature group has been used has recorded 20% of study and none of the study in this review had use five or six combination of features group. It can be concluded that, the importance of using more than one category of features cannot be neglected as 80 percent from the studies have used more than one group category features in forecasting students' performance whether in higher level institutions or high school.

Table 3 Groups of features used in the studies

Category Id	Features category description
C1	Demographic Features
C2	E-Learning Features
C3	Social Network Features
C4	School Design Features
C5	Academic Performance Features
C6	Previous Education Features

As tabulated in *Table 4*, there are different attributes that have been identify as the best predictors in predicting students' performance. The academic performance features have recorded the highest frequency of best attributes as illustrated in *Figure 2*. However, the findings differ from previous review article [11] that stated GPA is the most influential attribute to student performance prediction, as only two study that are [16, 43] out of 18 studies from the academic performance category that discovered that, and the majority of other studies ended up finding that students' grade and score in examinations, quizzes, and tests are the most impacted attributes to students' performance. This is because academic performance is mainly measured through the GPA, grades and score. As this review also focusing on the high school and secondary school level, the measurement to the students' performance not only depends on the GPA.

Previous education category feature has become the second most influential attributes as 8 of the studies found that Scholastic admission test score, high school marks and grade and National Examination score are able to make the highest impact on students' performance with the recurrence percentage is 47.5. Previous education features in predicting students' performance are essentially important as it represents benchmark of students' academic achievement [21]. It is commonly defined as score or grades obtained by students in the past level of education such as high school or university admission score which aids in understanding the consistency of students' performance.

As for the e-learning features, a total of 11 studies used the e-learning features combine with other category of features, and the findings have shown that 7 studies have recognized the most impacted attributes are the number of raise hands and logins, submission status for assignment and homework given, announcement view and students' participation in discussion or forums are impacted the accuracy of students' performance prediction. Furthermore, [19] identify that best variable in e-learning features are those that related to exercise and homework rather than students' participation in forum and discussions [17, 32]. For demographic features, although the recurrence percentage is the quite high among others that is 45 percent, however the best attribute is not as promising as the academic performance features. only 3 studies found the demographic features such as gender, caste, father and mother education, father and mother occupation, family income and family size do have greater impact to the prediction [16, 22, 34]. While [49] analysis has discovered that there is no corelation between students is first child or not in predictive model.

Although not many studies emphasize on using the school design features, however this review has able to identify two studies that has discovered that the best attributes in their research is school size [3] and the percentage of lecturer attendance [10]. Therefore, it is important for other researchers to look beyond the common attributes such as academic performance as the educators and school facilities also tend to impacted students' performance. While the social network features able to identify list of attributes such as List of webs visited; visits duration, time spend on movies online, time spend on reading online are able to contribute to the students' performance prediction [24, 30, 47].

Figure 4 illustrates the features that have been used in this review study. The features were divided based on six groups that have been discussed above. Some features are redundant due to their different labels. However, this study had identified those redundant features and relabeled them.

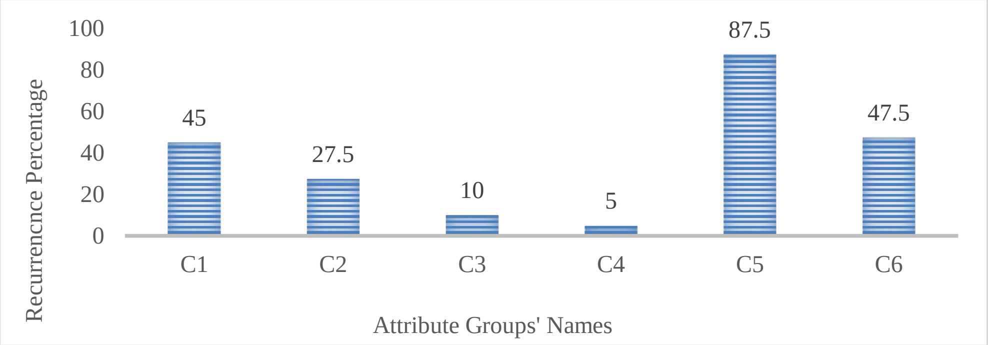

Figure 2 Most common used attributes in predicting students' performance based on groups category

Table 4 Best attributes in predicting students' performance

Features category	Best attributes	Methods	Paper Id
Academic performance features	Students' grade and score in examination, test, quizzes and assignment, GPA, internal assessment on courses and subjects, students' attendance marks.	Classification	P4,P6,P8,P9,P11,P14,P15,P16, P1,P20,P25,P26,P28,P32,P40
		Classification and Clustering	P19
		Regression	P26, P31
	Major course change	Classification	P30
Previous education features	National examination score, Scholastic admission test score, high school grades and score	Classification	P1,P2, P8, P10, P21, P23
		Regression	P2

Features category	Best attributes	Methods	Paper Id
		Clustering	P2, P34
Demographic Features	Gender, caste, father and mother education, fathers and mother occupation, family income	Classification	P4,P14,P22,P37
E-learning features	Number of raise hand, number of logins to the online class, announce view and participation in forum and discussions, submission of exercises and homework, hours spent on material.	Classification	P5, P7, P12, P13, P24
	Students' course marks and grade, attendance marks.	Classification and clustering	P29
		Classification	P18
Social Network Features	List of webs visited; visits duration, time spend on movies online, time spend on reading online.	Classification	P12, P18
		Clustering	P35
School Design Features	School size	Regression	P3
	Lecturer attendance percentage	Classification	P28

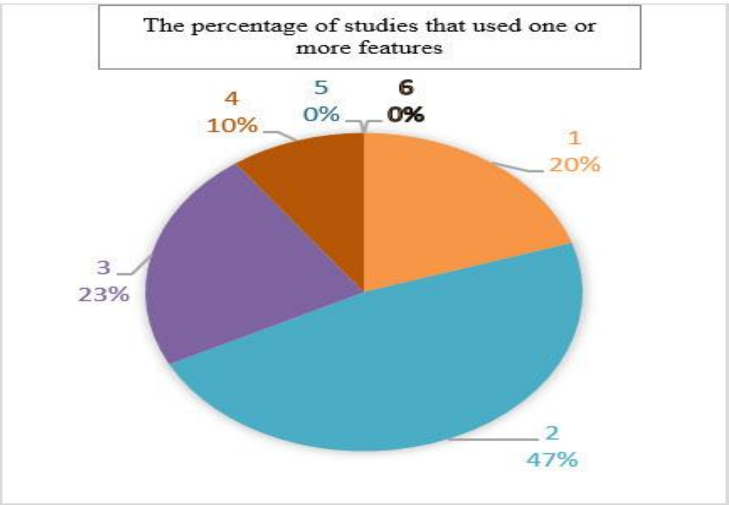

Figure 3 The percentage of studies using one or more number of features categories in their study

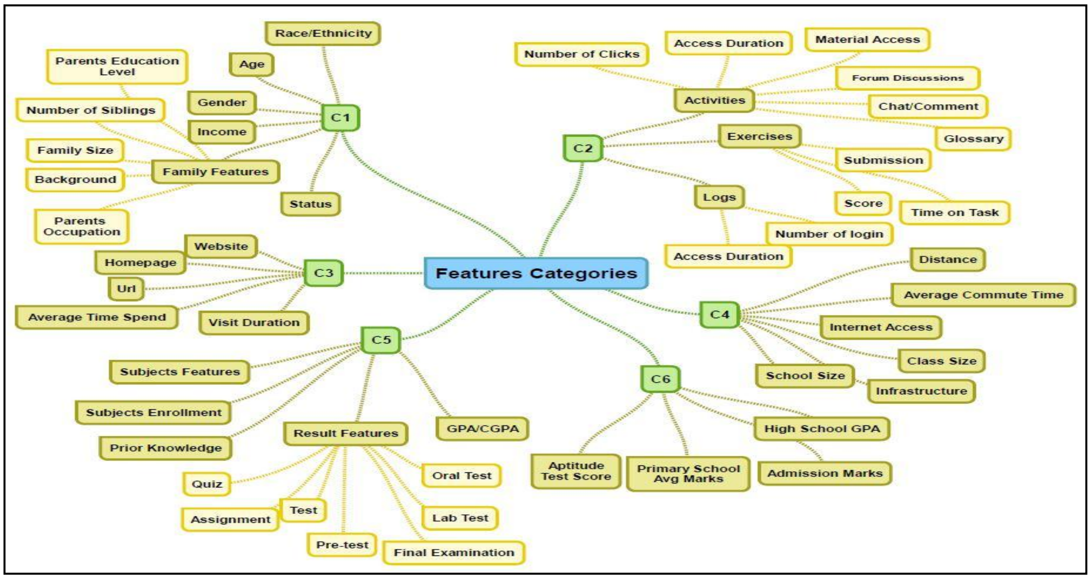

Figure 4 Some of features used in this study based on their groupings

3.2(RQ:ii) Which methods that are commonly used to investigate students' performance?

Figure 5 shows different researchers have applied various techniques such as classification, regression and clustering to predict students' performance. The goal of classification is to accurately predict the target class for each case in the data [54], while regressions are used to identify the relationship between dependent variables and independent variables [55]. Different from classification and regression, clustering is an unsupervised classification process that is used to group objects into classes of similar objects. The classification method is the most commonly used method in predicting student performance, as this review discovered that 30 out of 40 studies used the classification method rather than clustering and regression, each of which had only 5 studies. This is because the classification method or also known as supervised learning use labeled data to train the algorithms are computationally less complex compare to other methods. These labelled data also is used to train techniques by learn over time and finally accurately classify data or make predictions.

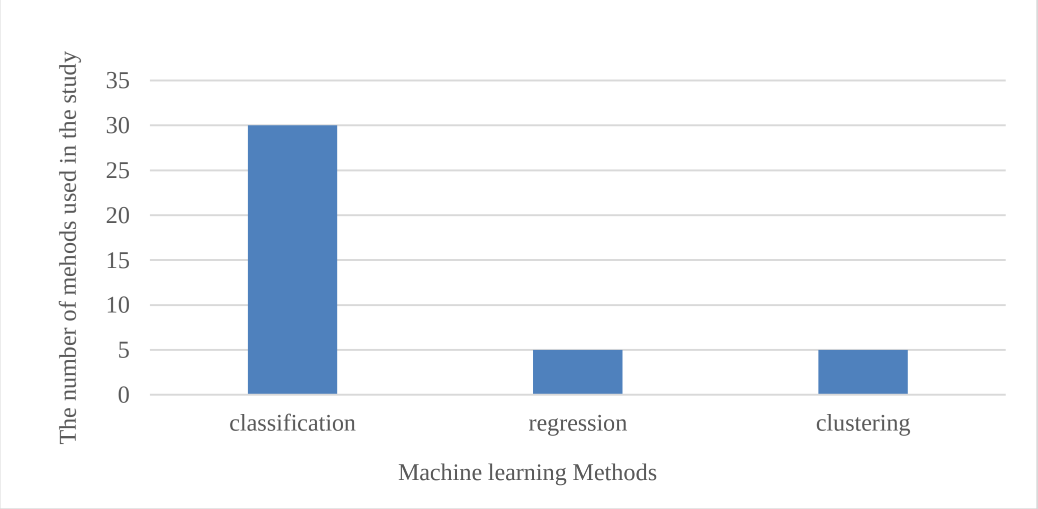

Figure 5 The number of students' performance prediction method in the review

3.3(RQ:iii) What are the best algorithms or techniques used in student performance prediction?

The techniques used in this review paper are numerous. This implies that there may be multiple options for implementing prediction algorithms. Furthermore, several models are often used in the same paper to make comparisons in order to find the best model suitable with their dataset. Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), Artificial Neural Network (ANN), Logistic Regression (LR), and K-Nearest Neighbour (KNN) are among the algorithms frequently used by researchers to predict students' performance. A brief explanation on results based on algorithms or techniques to predict students' performance will be discussed in the next section.

3.3.1Random forest (RF)

Random Forest (RF) is a supervised ensemble machine learning approach for classification, regression, and other tasks that operate by constructing a number of decision trees during the training time and producing the output of the class, which is the mode of the classes of the individual trees [56, 57]. This review identified 18 out of the 40 studies that had tested the Random Forest algorithm on their dataset for prediction. From the total of 18, 6 studies showed the Random Forest algorithm to have the highest accuracy beating other algorithms in the prediction of students at risk [17, 22, 24, 25] or students' dropout [43, 44]. Details of the feature categories and accuracy levels achieved according to paper ID are shown in *Table 5.*

Table 5 RF details

Paper Id	Features category	Accuracy
P5	C2	94%
P12	C1, C2, C6	71.6%
P13	C2, C5	84%
P14	C1, C5	99%
P31	C5	73%
P32	C1, C5	86.6%

3.3.2Support vector machine (SVM)

Social network analysts argue that causation is not located in individuals, but in the social structure.

Social network analysis examines the structure and composition of ties in a given network and provides insights into its structural characteristics [58]. In education, SVM algorithms have been proven to be helpful in monitoring students' interactions and participation in online courses. It has been acknowledged to be among the most reliable and accurate algorithms in most Machine Learning applications [54]. SVM comes in second for the number of highest accuracy rates achieved in this study. 5 over 19 papers used SVM to predict academic success [13], academic performance [19, 34, 45, 50] and students' pass rate [35]. For details on the accuracy rates achieved and feature categories used in these studies, refer *Table 6.*

Table 6 Support vector machine details

Paper Id	Features Category	Accuracy
P1	C1, C3, C5, C6	70.4%
P7	C5, C6	-
P22	C1, C5. C6	76.3%
P23	C1, C3, C5, C6	92.6%
P33	C1, C5	75.4%

3.3.3Artificial neural network (ANN)

Artificial Neural Network is a mathematical model inspired by biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. The network usually learns the connection weights from available training patterns. Performance is improved over time by iteratively updating the weights in the network [59]. From the total of 11 papers reviewed that used the ANN algorithm, 3 of them used ANN to forecast students' performance [48, 51, 52] and predict student course assessment course as tabular in *Table 7*. Recorded that ANN was able to achieve the highest accuracy [41].

Table 7 Artificial neural network details

Paper Id	Features category	Accuracy
P29	C2	80.4%
P36	C5	75%
P39	C5, C6	76
P40	C2	80%

3.3.4Logistic regression (LR) and linear regression (LNR)

Linear Regression predicts a continuous numeric output from a linear combination of attributes, while Logistic Regression predicts the odds of two or more outcomes, allowing for categorical predictions [8]. 4 among 11 papers were recorded in *Table 8* that have been used Logistic Regression techniques with their accuracy achieved. While *Table 9* shows the details of accuracy and features category used using Linear Regression algorithm.

Table 8 Logistic regression details

Paper Id	Features category	Accuracy
P4	C1, C4, C5, C6	67%
P11	C5, C6	51.9%
P15	C5	89.15%
P24	C1, C5	88.8%

Table 9 Linear regression details

Paper Id	Features category	Accuracy
P20	C1, C2, C5	50%
P21	C4, C6	60.24%
P30	C1, C5.C6	-

3.3.5Decision tree

The Decision Tree classification technique is performed in two stages, which are 1) tree building, and 2) pruning [60]. The internal nodes of the tree represent conditions, the external nodes or the leaves represent class labels, while branches from the internal nodes represent outcomes of the tests or conditions [61]. Decision Tree (DT) is one of the famous algorithms used for predictive modelling on educational data [62]. 19 studies reviewed used the Decision Tree algorithm, and 3 of them succeeded to achieve the highest rate of accuracy when competing with other algorithms. For example, [26] used academic performance and previous education features to predict students' performance in intermediate and secondary schools. Decision Tree also gave good accuracy in identifying high-risk students who need timely help to complete their studies as discovered by [30] who used the combination of e-learning, social network, and academic performance features to conduct their work.

Table 10 Decision tree details

Paper Id	Features category	Accuracy
P10	C5, C6	96.6%
P18	C2, C3, C5	91.9%
P37	C1, C5	94%

3.3.6Naïve bayes (NB)

The Naive Bayes classifier simplifies learning by assuming that features are independent of given class and provide probabilistic interpretations of classifications [63]. Although independence is generally a poor assumption, in practice, Naive Bayes often competes well with other sophisticated classifiers. [18] identified that the Naïve Bayes algorithm has better accuracy in predicting the performance of junior high school students, while

[21] used Naïve Bayes to analyse undergraduate students' performance.

Table 11 Naïve bayes details

Paper Id	Features category	Accuracy
P6	C5	69%
P9	C5, C6	83.65%

3.3.7Hybrid algorithms

In prediction models, the challenging task is to choose the effective techniques that could produce satisfying predictive accuracy [64]. Some researchers introduced the hybrid approach of combining a few machine learning algorithms together to achieve maximum accuracy. Hybrid approaches are defined as incorporating a number of possible machine learning algorithms in order to achieve better performance than any examined single learning algorithms [65]. In this review, five papers used hybrid algorithm approaches to predict student performance [15, 38, 20, 50], and to analyse students at risk in a course [28]. Details of the algorithm integrations are illustrated in *Table 12*.

Table 12 Hybrid algorithms details

Paper Id	Features category	Accuracy
P3	C1, C4, C5, C6	-
P8	C1, C6	82.3%
P16	C5	85%
P26	C1, C5	98.96%
P38	C1, C5	97.93

Other than the algorithms mentioned above, the ID3, K-Nearest Neighbour (KNN), Gradient Boosting (GB), Association Rule Mining (ARM), J48, RMSR, and Fuzzy Network (FN) algorithms were found to have achieved high accuracy rates in only one paper reviewed. Details on the accuracy rates based on the different techniques used are tabulated in *Table 13*.

Table 13 Other algorithms details

Paper Id	Features category	Algorithms	Accuracy
P2	C1, C5, C6	RMSR	94%
P17	C5, C6	ID3	80%
P25	C5, C6	FN	80%
P28	C2, C5	KNN	89%
P27	C5	ARM	67.3%
P34	C1, C5, C6	GBoost	87.9%
P35	C2, C5, C6	J48	94.7%

Details of the algorithms frequency in terms of as a best predictor or frequency used for comparison in order to find more accurate results in each study are illustrated in *Figure 6* below. *Figure 6* shows that Decision Tree is the most commonly used algorithm in the study, with 19 out of 40 studies using it to determine the best accuracy for educational datasets. However, Random Forest placed first in terms of highest percentage obtained because it provided the highest accuracy in 6 over 18 studies or a total of 33.5% when compared to others. When it comes to the highest accuracy that each algorithm can achieve, Random Forest once again demonstrated the best performance with 99% accuracy results [26], putting this algorithm in first place among others. However, SVM also performs as good as Random Forest as 19 papers have tested SVM as a single or comparison algorithms and SVM able to gain 5 over 19 best accuracies compare to Random Forest. While other algorithms that obtained high frequency in the chosen algorithms are ANN, NB, KNN and LR.

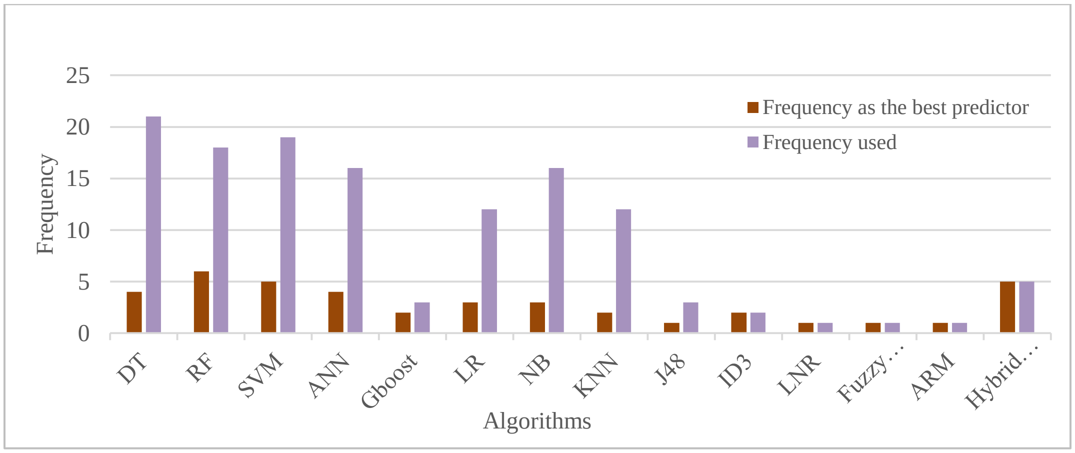

Figure 6 The Algorithms frequency details as the best predictor or used as a comparisons technique in finding the best accuracy

4.Conclusion

Performance prediction has evolved into a useful research topic that assists educators, academics, policymakers, and management in improving the teaching and learning process. This paper presents a 5-year systematic review of attributes used in student performance prediction made via machine learning, educational data mining, and learning analytics approaches, as well as their applicability in the context of student performance prediction. Analysis on 40 papers included in this review brought about great discussions on the varieties of features and algorithms that can impact student performance prediction.

The majority of features used in predicting students' performance are from the academic feature category, which includes students' grade and score in examinations, tests, quizzes, and assignments, GPA, internal assessment on courses and subjects, and attendance marks. According to the findings of this review, academic performance features able to outperform other categories of features in terms of the best attributes for students' performance prediction. Another finding from this review is that the examination marks and score are the best attributes compare to GPA as highlighted by [11]. This is due to the fact that prediction of student performance has been widely adopted in the educational sector, not only at the higher levels of educational attainment, but also in high school, middle school, and secondary school. This will be a strong foundation for educators and administrators to be able to monitor their students' academic progress while also establishing a suitable approach based on their students' strengths.

Looking deeper into the frequency of techniques used by researchers in predicting student performance, Decision Tree, Random Forest, SVM, ANN, and NB show a strong competition of most common used techniques. However, among the 40 papers reviewed, the Random Forest algorithm had the best performance at 99 percent when using demographic and academic performance features [26], as well as the majority frequency in terms of highest accuracy in student performance prediction. RF could be used for both regression and classification method and even though it is handling high-dimensional data, the lesser time for processing make it better than Decision Tree algorithm [51].

The hybrid approach which combined three algorithms: AODE, IBK, and J48, scored the second highest prediction accuracy at 98.96%. Previous research on student performance prediction suggest that algorithm combinations can help gain better accuracy, but they do not emphasize on the importance of incorporating algorithms to improve accuracy results. Not only has the integration of multiple features resulted in improved accuracy prediction, but the combination of different techniques has also contributed to improved accuracy results [66]. These results show that the selection of combined features categories together with suitable algorithms can affect the accuracy levels obtained when analysing students' performance. However, the significance of using hybrid algorithms in student performance prediction requires further investigation as most hybrid approach has the capacity to produce competitive performance when compared with related methods. Finally, it is hoped that academic forecasting research on the educational system can help students in improving their academic performance, and educators to understand their students' needs. Additionally, the findings can help the educational management to design more efficient curricula for better education adaptation. A complete list of abbreviations is shown in *Appendix I.*

Limitations

Since there is no evidence in favour of smaller classes, most articles are concerned about overfitting. As a result, the use of data sampling in the application of data mining to educational datasets should indeed be outlined in order to overcome the overfitting and underfitting problem. Other issues that need to be highlighted in this review, there are too limited papers that study on the school design and social network features as mostly researchers are emphasizing on the academic performance and demographic features. Hopefully, more researchers will implement more categories of features in the future in order to find the best attributes suitable for adaptation to specific algorithms.

Acknowledgment

This research was supported by the Research Management and Innovation Centre, Universiti Sultan Zainal Abidin (UniSZA).

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Gaftandzhieva S, Docheva M, Doneva R. A comprehensive approach to learning analytics in Bulgarian school education. Education and Information Technologies. 2021; 26(1):145-63.

[2] Alyahyan E, Düştegör D. Predicting academic success in higher education: literature review and best practices. International Journal of Educational Technology in Higher Education. 2020; 17(1):1-21.

[3] Ferreira SA, Andrade A. Academic analytics: anatomy of an exploratory essay. Education and Information Technologies. 2016; 21(1):229-43.

[4] Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. Journal of Big Data. 2015; 2(1):1-21.

[5] Khatib KC, Kamble TD, Chendake BR, Sonavane GN. Social media data mining for sentiment analysis. International Research Journal of Engineering and Technology. 2016; 3(4):373-6.

[6] Ozdemir D, Opseth HM, Taylor H. Leveraging learning analytics for student reflection and course evaluation. Journal of Applied Research in Higher Education. 2019; 12(1):27-37.

[7] Nuutila K, Tuominen H, Tapola A, Vainikainen MP, Niemivirta M. Consistency, longitudinal stability, and predictions of elementary school students' task interest, success expectancy, and performance in mathematics. Learning and Instruction. 2018; 56:73-83.

[8] Lang C, Siemens G, Wise A, Gasevic D. Handbook of learning analytics. New York, NY, USA: SOLAR, Society for Learning Analytics and Research; 2017.

[9] Nawang H, Makhtar M, Shamsudin SN. Classification model and analysis on students performance. Journal of Fundamental and Applied Sciences. 2017; 9(6S):869-85.

[10] Tsai YS, Gasevic D. Learning analytics in higher education-challenges and policies: a review of eight learning analytics policies. In proceedings of the seventh international learning analytics & knowledge conference 2017 (pp. 233-42).

[11] Shahiri AM, Husain W. A review on predicting student's performance using data mining techniques. Procedia Computer Science. 2015; 72:414-22.

[12] Kitchenham B, Brereton OP, Budgen D, Turner M, Bailey J, Linkman S. Systematic literature reviews in software engineering–a systematic literature review. Information and Software Technology. 2009; 51(1):7-15.

[13] Gil PD, Da CMS, Moro S, Costa JM. A data-driven approach to predict first-year students' academic success in higher education institutions. Education and Information Technologies. 2021; 26(2):2165-90.

[14] Qazdar A, Er-raha B, Cherkaoui C, Mammass D. A machine learning algorithm framework for predicting students performance: a case study of baccalaureate students in Morocco. Education and Information Technologies. 2019; 24(6):3577-89.

[15] Costa-mendes R, Oliveira T, Castelli M, Cruz-jesus F. A machine learning approximation of the 2015 Portuguese high school student grades:a hybrid approach. Education and Information Technologies. 2021; 26(2):1527-47.

[16] Baars GJ, Stijnen T, Splinter TA. A model to predict student failure in the first year of the undergraduate medical curriculum. Health Professions Education. 2017; 3(1):5-14.

[17] Youssef M, Mohammed S, Hamada EK, Wafaa BF. A predictive approach based on efficient feature selection and learning algorithms' competition: case of learners' dropout in MOOCs. Education and Information Technologies. 2019; 24(6):3591-618.

[18] Kostopoulos G, Kotsiantis S, Verykios VS. A prognosis of junior high school students' performance based on active learning methods. In international conference on brain function assessment in learning 2017 (pp. 67-76). Springer, Cham.

[19] Moreno-marcos PM, Pong TC, Munoz-merino PJ, Kloos CD. Analysis of the factors influencing learners' performance prediction with learning analytics. IEEE Access. 2020; 8:5264-82.

[20] Al-obeidat F, Tubaishat A, Dillon A, Shah B. Analyzing students' performance using multi-criteria classification. Cluster Computing. 2018; 21(1):623-32.

[21] Asif R, Merceron A, Ali SA, Haider NG. Analyzing undergraduate students' performance using educational data mining. Computers & Education. 2017; 113:177-94.

[22] Yousafzai BK, Hayat M, Afzal S. Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. Education and Information Technologies. 2020; 25(6):4677-97.

[23] Adekitan AI, Noma-osaghae E. Data mining approach to predicting the performance of first year student in a university using the admission requirements. Education and Information Technologies. 2019; 24(2):1527-43.

[24] Azcona D, Hsiao IH, Smeaton AF. Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints. User Modeling and User-Adapted Interaction. 2019; 29(4):759-88.

[25] Akçapınar G, Hasnine MN, Majumdar R, Flanagan B, Ogata H. Developing an early-warning system for spotting at-risk students by using eBook interaction logs. Smart Learning Environments. 2019; 6(1):1-15.

[26] Hussain S, Dahan NA, Ba-alwib FM, Ribata N. Educational data mining and analysis of students' academic performance using WEKA. Indonesian Journal of Electrical Engineering and Computer Science. 2018; 9(2):447-59.

[27] Adekitan AI, Salau O. The impact of engineering students' performance in the first three years on their graduation result using educational data mining. Heliyon. 2019; 5(2):1-21.

[28] Marbouti F, Diefes-dux HA, Madhavan K. Models for early prediction of at-risk students in a course using standards-based grading. Computers & Education. 2016; 103:1-15.

[29] Altujjar Y, Altamimi W, Al-turaiki I, Al-razgan M. Predicting critical courses affecting students

performance: a case study. Procedia Computer Science. 2016; 82:65-71.

[30] Zhou Q, Quan W, Zhong Y, Xiao W, Mou C, Wang Y. Predicting high-risk students using internet access logs. Knowledge and Information Systems. 2018; 55(2):393-413.

[31] Aydoğdu Ş. Predicting student final performance using artificial neural networks in online learning environments. Education and Information Technologies. 2020; 25(3):1913-27.

[32] Karlos S, Kostopoulos G, Kotsiantis S. Predicting and interpreting students' grades in distance higher education through a semi-regression method. Applied Sciences. 2020; 10(23):1-19.

[33] Iqbal MS, Luo B. Prediction of educational institution using predictive analytic techniques. Education and Information Technologies. 2019; 24(2):1469-83.

[34] Zohair LM. Prediction of student's performance by modelling small dataset size. International Journal of Educational Technology in Higher Education. 2019; 16(1):1-8.

[35] Ma X, Zhou Z. Student pass rates prediction using optimized support vector machine and decision tree. In 8th annual computing and communication workshop and conference (CCWC) 2018 (pp. 209-15). IEEE.

[36] Hashim AS, Awadh WA, Hamoud AK. Student performance prediction model based on supervised machine learning algorithms. In IOP conference series: materials science and engineering 2020 (pp. 1-19). IOP Publishing.

[37] Hamsa H, Indiradevi S, Kizhakkethottam JJ. Student academic performance prediction model using decision tree and fuzzy genetic algorithm. Procedia Technology. 2016; 25:326-32.

[38] Pandey M, Taruna S. Towards the integration of multiple classifier pertaining to the student's performance prediction. Perspectives in Science. 2016; 8:364-6.

[39] Badr G, Algobail A, Almutairi H, Almutery M. Predicting students' performance in university courses: a case study and tool in KSU mathematics department. Procedia Computer Science. 2016; 82:80-9.

[40] Akçapınar G, Altun A, Aşkar P. Using learning analytics to develop early-warning system for at-risk students. International Journal of Educational Technology in Higher Education. 2019; 16(1):1-20.

[41] Hussain M, Zhu W, Zhang W, Abidi SM, Ali S. Using machine learning to predict student difficulties from learning session data. Artificial Intelligence Review. 2019; 52(1):381-407.

[42] Zheng G, Fancsali SE, Ritter S, Berman S. Using instruction-embedded formative assessment to predict state summative test scores and achievement levels in mathematics. Journal of Learning Analytics. 2019; 6(2):153-74.

[43] Rovira S, Puertas E, Igual L. Data-driven system to predict academic grades and dropout. PLoS One. 2017; 12(2).

[44] Rodríguez-muñiz LJ, Bernardo AB, Esteban M, Díaz I. Dropout and transfer paths: what are the risky profiles when analyzing university persistence with machine learning techniques?. Plos One. 2019; 14(6):1-21.

[45] Francis BK, Babu SS. Predicting academic performance of students using a hybrid data mining approach. Journal of Medical Systems. 2019; 43(6):1-15.

[46] Aiken JM, De BR, Hjorth-jensen M, Caballero MD. Predicting time to graduation at a large enrollment American university. Plos One. 2020; 15(11).

[47] Hussain M, Zhu W, Zhang W, Abidi SM. Student engagement predictions in an e-learning system and their impact on student course assessment scores. Computational Intelligence and Neuroscience. 2018:1-21.

[48] Czibula G, Mihai A, Crivei LM. S PRAR: a novel relational association rule mining classification model applied for academic performance prediction. Procedia Computer Science. 2019; 159:20-9.

[49] Matzavela V, Alepis E. Decision tree learning through a predictive model for student academic performance in intelligent M-learning environments. Computers and Education: Artificial Intelligence. 2021.

[50] Viloria A, López JR, Leyva DM, Vargas-mercado C, Hernández-palma H, Llinas NO, et al. Data mining techniques and multivariate analysis to discover patterns in university final researches. Procedia Computer Science. 2019; 155:581-6.

[51] Deng H, Wang X, Guo Z, Decker A, Duan X, Wang C, et al. Performancevis: visual analytics of student performance data from an introductory chemistry course. Visual Informatics. 2019; 3(4):166-76.

[52] Çetinkaya A, Baykan ÖK. Prediction of middle school students' programming talent using artificial neural networks. Engineering Science and Technology, an International Journal. 2020; 23(6):1301-7.

[53] Mokhairi M, Nawang H, Wan SN. Analysis on students performance using naïve. Journal of Theoretical and Applied Information Technology. 2017; 31(16):3993-4000.

[54] Hu H, Zhang G, Gao W, Wang M. Big data analytics for MOOC video watching behavior based on spark. Neural Computing and Applications. 2020; 32(11):6481-9.

[55] Slater S, Joksimović S, Kovanovic V, Baker RS, Gasevic D. Tools for educational data mining: a review. Journal of Educational and Behavioral Statistics. 2017; 42(1):85-106.

[56] Breiman L. Random forests. Machine Learning. 2001; 45(1):5-32.

[57] Yusuf A. Prediction of students'performance in E-learning environment using random forest. Doctoral Dissertation, University of Technology Malaysia.

[58] Noble WS. What is a support vector machine?. Nature Biotechnology. 2006; 24(12):1565-7.

[59] Gupta N. Artificial neural network. Network and Complex Systems. 2013; 3(1):24-8.

[60] Hamoud A, Hashim AS, Awadh WA. Predicting student performance in higher education institutions using decision tree analysis. International Journal of Interactive Multimedia and Artificial Intelligence. 2018; 5:26-31.

[61] Zulfiker MS, Kabir N, Biswas AA, Chakraborty P, Rahman MM. Predicting students' performance of the private universities of Bangladesh using machine learning approaches. International Journal of Advanced Computer Science and Applications. 2020; 11(3):672-9.

[62] Sivakumar S, Venkataraman S, Selvaraj R. Predictive modeling of student dropout indicators in educational data mining using improved decision tree. Indian Journal of Science and Technology. 2016; 9(4):1-5.

[63] Rish I. An empirical study of the naive Bayes classifier. In IJCAI workshop on empirical methods in artificial intelligence 2001 (pp. 41-6).

[64] Sokkhey P, Okazaki T. Hybrid machine learning algorithms for predicting academic performance. International Journal of Advanced Computer Science and Applications. 2020; 11(1):32-41.

[65] Dole L, Rajurkar J. A decision support system for predicting student performance. International Journal of Innovative Research in Computer and Communication Engineering. 2014; 2(12):7232-7.

[66] Mohamad M, Makhtar M, Abd RMN. The reconstructed heterogeneity to enhance ensemble neural network for large data. In international conference on soft computing and data mining 2016 (pp. 447-55). Springer, Cham

Hasnah Nawang earned a BSc in Computer Science from Universiti Putra Malaysia in 2006 and an MSc in Computer Science from Universiti Sultan Zainal Abidin in Terengganu, Malaysia, in 2018. She is currently a PhD candidate at the Department of Computer Science in the Faculty of Computing and Informatics at Universiti Sultan Zainal Abidin in Terengganu, Malaysia. Since 2007, she has also worked as a secondary school teacher in the Department of Mathematics and Computer Science. Machine Learning, Data Mining, and Deep Learning are among her current research interests.
Email: hasnah.nawang@gmail.com

Dr. Mokhairi Makhtar received his PhD from University of Bradford, United Kingdom in 2012. He is currently a Professor in the Department of Computer Science, Universiti Sultan Zainal Abidin, Terengganu, Malaysia. His current research interests include Machine Learning, Ensemble Method, Data Mining, Soft Computing, Timetabling and Optimisation, Natural Languange Processing, E-Learning and Deep Learning.
Email: mokhairi@unisza.edu.my

Dr. Wan Mohd Amir Fazamin Wan Hamzah received his PhD from Universiti Malaysia Terengganu, Malaysia. He is currently a lecturer in Universiti Sultan Zainal Abidin, Terengganu, Malaysia. His research interests include Learning Analytics, Gamification, e-Learning and Cloud Computing.
Email: amirfazamin@unisza.edu.my

Appendix I

S. No.	Abbreviations	Descriptions
1	ANN	Artificial Neural Network
2	ARM	Association Rule Mining
3	DM	Data Mining
4	DT	Decision Tree
5	FN	Fuzzy Network
6	GBoost	Gradient Boosting
7	ID3	Iterative Dichotomiser 3
8	J48	J48 algorithm
9	KNN	Nearest Neighbor
10	LA	Learning Analytics
11	LNR	Linear Regression
12	LR	Logistic Regression
13	NB	Naïve Bayes
14	ML	Machine Learning
15	RF	Random Forest
16	SVM	Support Vector Machine