# Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021)

https://doi.org/10.3991/ijet.v17i05.27685

Muhammad Haziq bin Roslan, Chwen Jen Chen(✉)
Universiti Malaysia Sarawak, Kota Samarahan, Malaysia
cjchen@unimas.my

**Abstrac**t—This systematic literature review aims to identify the recent research trend, most studied factors, and methods used to predict student academic performance from 2015 to 2021. The PRISMA framework guides the study. The study reviews 58 out of 219 research articles from Lens and Scopus databases. The findings indicate that the research focus of current studies revolves around identifying factors influencing student performance, data mining (DM) algorithms performance, and DM related to e-Learning systems. It also reveals that student academic records and demographics are primary aspects that affect student performance. The most used DM approach is classification and the Decision Tree classifier is the most employed DM algorithm.

**Keywords—**Educational Data Mining (EDM), data mining (DM) techniques, prediction studies, student academic performance, systematic literature review

## 1 Introduction

Student academic performance is a significant aspect in determining educational success at all levels [1, 2, 3]. Academic performance is crucial for students to continue their studies and secure their future [4, 5]. Several studies uncover factors that can predict their performance. Earlier studies such as McKenzie and Schweitzer [6], Andujar et al. [7], and Taniguchi et al. [8] employed classical statistical methods to determine these factors. Generally, regression analysis, discriminant analysis, and cluster analysis are examples of classical statistics approaches used in this area [9]. Artificial intelligence methods such as Backpropagation, Support Vector Regression, Gradient Boosting Classifier [10], Bayesian Classifier, Artificial Neural Network, and Decision Tree [11] are later employed. The latter involves a mix of advanced statistical methods and artificial intelligence heuristics and has contributed to the growth of educational data mining (EDM), which adds to our understanding of how to predict student academic performance [12, 13, 14, 15, 16]. EDM research studies [17] applied data mining (DM) techniques to data obtained from diverse educational systems to improve the quality of education [18]. Such mining is significant as it enables educators to take necessary interventions to achieve optimal student performance [19].

To date, scholars are debating how to anticipate students' academic performance, focusing on the types of variables that influence such a prediction [20, 21]. Based on a

SLR on EDM research by Saa et al. [22], several factors that predict student performance and learning outcomes are identified. These factors are divided into four main categories which are student internal and external assessments, student e-Learning activity, student demographics as well as social information. In a SLR on EDM studies, Shahiri et al. [23] discovered that the cumulative grade point average (CGPA), as well as the inner assessment, are widely employed to predict student performance. Other key characteristics mentioned by Shahiri et al. [23] include demographics, external assessment, extracurricular participation, high school background, and social interaction network. These findings are in line with Alyahyan and Düştegör's recent study [24]. Alyahyan and Düştegör [24] reveal that students' past academic qualifications, demographics, e-Learning activities, psychological aspects, and environment are the commonly reported factors.

Various techniques are utilized to predict students' performance and the analysis by Peña-Ayala [25] has revealed that 60% of past EDM studies utilized predictive approaches while the remaining 40% employed descriptive approaches. Furthermore, classification and clustering are the most popular methodologies employed in EDM studies, according to a review by Saa et al. [22]. Decision Tree, Neural Network, Bayesian Networks, Rule Induction, and Support Vector Machines are examples of classification techniques that are frequently used to perform prediction [24]. Among these DM techniques, Neural Network is reported to provide the highest prediction accuracy, followed by Decision Tree [23]. Other research, such as those by Devasia et al. [26] and Rifat et al. [27], have found that the algorithms' accuracy varies. Hence, a thorough analysis of past related studies will help to unveil techniques that are often used and can produce good predictions.

Over the past years, several SLRs related to the prediction of student performance were reported. Shahiri et al. [23] did a SLR on EDM studies to predict student performance from 2002 until early 2015. Another review by Saa et al. [22] was based on 36 EDM studies from 2009 to 2018 and the research uncovered characteristics that influence student performance in higher education. Both reviews are concerned with finding aspects that influence student performance as well as the DM techniques used. Alyahyan and Düştegör [24] reviewed 17 papers from 2015 to 2019 that were connected to anticipate academic performance in tertiary education. However, this study aims to build a set of criteria for educators to employ when using DM approaches to predict student performance. Another review study by Dutt et al. [28] emphasizes clustering algorithms employed over three decades, from 1983 to 2016. Nevertheless, this review is limited to the clustering approach.

The rapid advancement in DM techniques that offer promising educational insights as well as the proliferation of EDM studies for educational predictions in recent years, points to the need to expand existing SLR endeavors to compile a more comprehensive list of factors that affect student performance and the DM techniques employed to predict this performance.

SLR can be defined as a structured way to collect, critically analyze, incorporate, and present findings on a research issue or subject of interest from multiple research

studies [29]. SLR has more benefits compared to traditional literature review [30] Although, the conventional review is generally faster and easier to conduct SLR is considered more reliable and unbiased [31].

This SLR study attempts to examine prediction-focus EDM studies that were published in the Scopus and Lens databases between 2015 and June 2021 to identify the (a) research trend (b) most studied factors that affect student performance and (c) DM methods employed.

## 2 Methodology

The study used the SLR approach based on guidelines by Kitchenham et al. [32]. Planning, conducting, and reporting were the three key phases. The following two subsections cover the first two phases, the planning phase and the conducting phase, while Section 3 covers the reporting phase.

### 2.1 Planning

In SLR, the initial step is planning. It consists of five phases, as detailed in the subsections below.

**Determine research questions.** In developing research questions, this study used the criteria provided by Kitchenham et al. [32]. Table 1 shows the features and details that drive the formation of research questions.

**Table 1.** Criteria of research questions

| Criteria | Details |
|---|---|
| Population/Participant | School students, university students (student performance) |
| Intervention | Prediction techniques |
| Outcome | Research trend, factors affecting student performance, DM techniques used |
| Context | Educational institutions |

Based on the details in Table 1, the following are the research questions formed:

- RQ1: What is the research trend of prediction-focus EDM studies?
- RQ2: What are the aspects that affect student performance?
- RQ3: What are the DM approaches used to predict student performance?

**Identify keywords.** The search keywords were derived based on the research questions stated. In this study, the search string used is: [("educational data mining" OR "data mining") AND ("predicting student performance" OR "predicting student achievement OR "factors affecting student performance" OR "factors affecting student achievement")].

**Identify the source.** Lens and Scopus were the online databases chosen for this research. The Lens platform aims to aid institutional problem solving by sourcing, merging, and linking open knowledge sets, including comprehensive scholarly works

(https://www.lens.org/). The Scopus database contains comprehensive, expertly curated content across a wide variety of disciplines (https://www.scopus.com/).

**Determine the criteria for inclusion and exclusion.** Table 2 lists the criteria for inclusion and exclusion. This SLR included all articles retrieved from the search results, which met the given criteria.

**Table 2.** Criteria for inclusion and exclusion

| Inclusion criteria | Exclusion criteria |
|---|---|
| Journal articles | Conference papers, review papers, books, magazines, editorials, notes, and short survey |
| Primary literature | Secondary literature, tertiary literature |
| Published between 2015 and June 2021 | Published before 2015 and after June 2021 |
| Meet the research keywords | Duplicate reports of the same database |
| Classified as EDM studies | |
| Must be full-text, available, and accessible article | |
| Must be written in English language | |

**Choose a data extraction strategy.** The next step is to collect data from each study by searching the databases using the search string previously defined. Table 3 lists and describes the fields of the data collected. This study removed journal papers that did not provide appropriate information for one or more fields.

**Table 3.** Data layout

| Item | Description |
|---|---|
| Article ID | An ID number is given to each article so that the article can be accessed easily during the review |
| Article title | The title of the journal article |
| Author | The author(s) of the journal article |
| Year | The publication year of the journal article |
| Country | The country in which the research was conducted |
| Factors that influence student performance | Factors affecting student performance that were studied, such as student demographics, psychological aspects, and so forth |
| DM approaches | DM approaches used in the research, such as classification, clustering, regression, and so forth |
| DM techniques | DM algorithms used in the research, such as Support Vector Machine, Neural Network, Decision Tree, and so forth |
| Context | Participants of the study |

## 2.2 Conducting the review

The second phase of this SLR approach focuses on conducting the review. The four sections of this phase are as shown below.

**Identify the research.** In this phase, the search string as determined in Section 2.1 was used to search the Lens and Scopus online databases, respectively. Table 4 displays

the initial search results of the search engines. The Lens database retrieved 157 results, and the Scopus database retrieved 93 results.

**Table 4.** Online databases

| Online Database | Results |
|---|---|
| Lens | 157 |
| Scopus | 93 |

**Selecting the studies.** The literature research employed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines [33]. Figure 1 presents the PRISMA flowchart. This SLR only considered relevant journal papers that satisfied the inclusion criteria and examined the content of each chosen article to ensure its eligibility for inclusion.

The cross-checking of search results from both databases revealed 30 similar articles. Hence, this study excluded these duplicated articles. Then, the articles were screened and 20 of them were excluded for not meeting the inclusion criteria or not related to the focus of this study. The screening that involved a semi-automatic paper selection through examination of the whole text of the remaining articles also discovered that 90 papers were not DM studies. Moreover, 24 publications did not provide the full-text, 13 others were review articles, and 15 articles were not related to the focus of this study. This study further excluded these articles. Hence, this SLR only considered 58 journal papers. Table 6 (See "Appendix") shows all selected articles.

**Extracting the data.** Table 3 shows the data layout used to extract the information of selected articles. Such information served as data that helped the researcher gain significant insights to address the research questions of this study. Table 7 (See "Appendix") shows the data extracted for the selected articles.

**Synthesizing the data.** From the 58 selected articles, this SLR extracted three categories of research focus. These include identifying factors influencing student performance, DM algorithms performance, and DM related to e-Learning systems. Table 8 shows the research foci of the selected articles (See "Appendix"). Next, this study identified nine distinct categories of factors that influence student performance. Table 9 shows the categories of factors that affect student performance and their descriptions (See "Appendix"). Based on the 89 DM algorithms used to predict student performance, this study also identified three categories of DM approaches. Table 10 indicates the categories of DM approaches for the selected articles (See "Appendix").

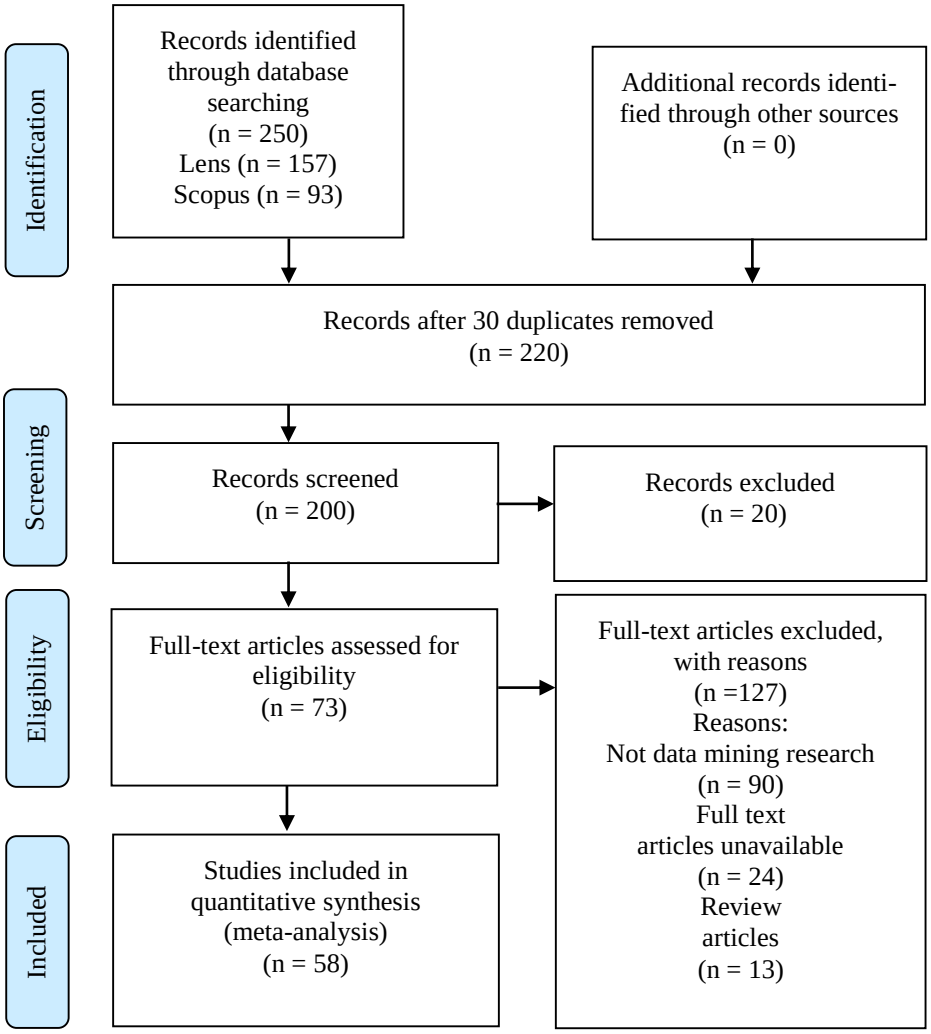


**Fig. 1.** PRISMA flowchart

## 3 Result

The third phase of SLR is reporting. This section summarizes and reports on the analysis of the selected articles according to research questions.

### 3.1 RQ1: What is the research trend of prediction-focus EDM studies?

This study analyzed the research trend according to four aspects: research focus, publication year, first author's country of origin, and study context.

**Research focus.** The research focus was classified into three main categories. Figure 2 indicates the three main categories of research focus. Identifying factors influencing student performance is the most researched, accounting for 49%, followed by DM algorithms performance (42%), and finally, DM to related e-Learning systems (9%).

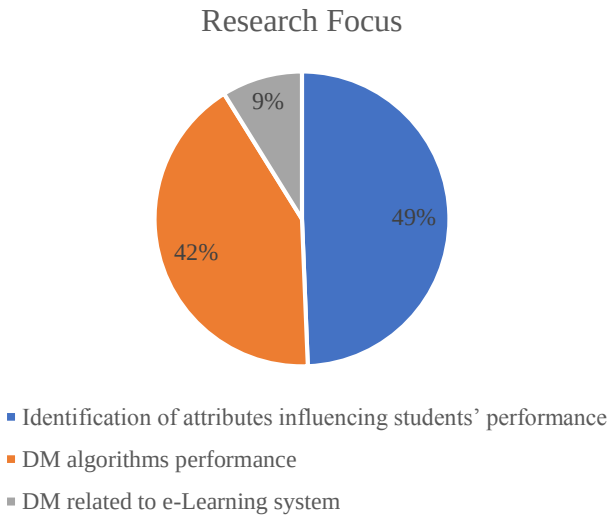


**Fig. 2.** Research focus

**Publication year.** Figure 3 depicts the number of articles published within the chosen period. The line graph is going up, which indicates studies in this context are on an upward trend. The highest number of articles, 19 altogether, were published in 2019. In the following year, there was a modest decline in the number of articles published, in which there were only 15. In 2015 and 2016, the number of articles published was the same, which is four. In 2017 and 2018, six and eight articles were published, respectively. As of June 2021, only two articles were published.

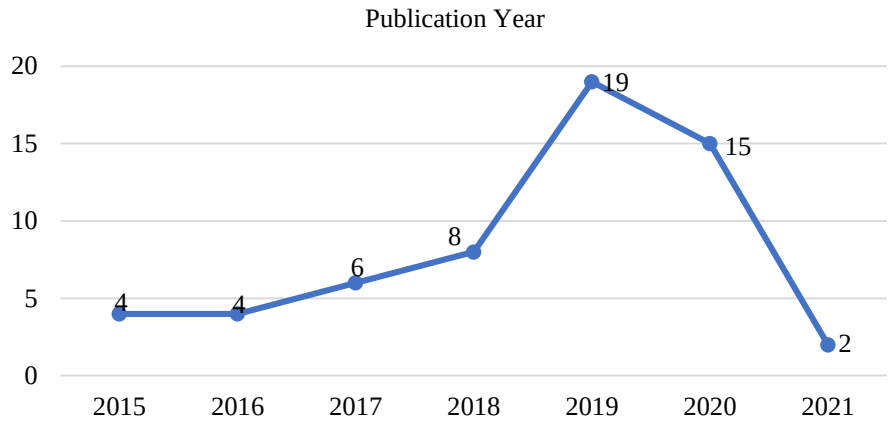


**Fig. 3.** Publication year of the selected research (2015 until June 2021)

**First author's country of origin.** Figure 4 shows the countries of origin of all first authors. With nine articles each, India and United Kingdom have the most published articles, followed by China with eight articles. Morocco, Philippines, and the United States of America tying for third place with three articles each. Cambodia, Canada, Iraq, Malaysia, Nigeria, and Spain have respectively produced two articles. Only one article was published in Egypt, Estonia, Indonesia, Ireland, Israel, Oman, Portugal, South Africa, Thailand, and Vietnam respectively.

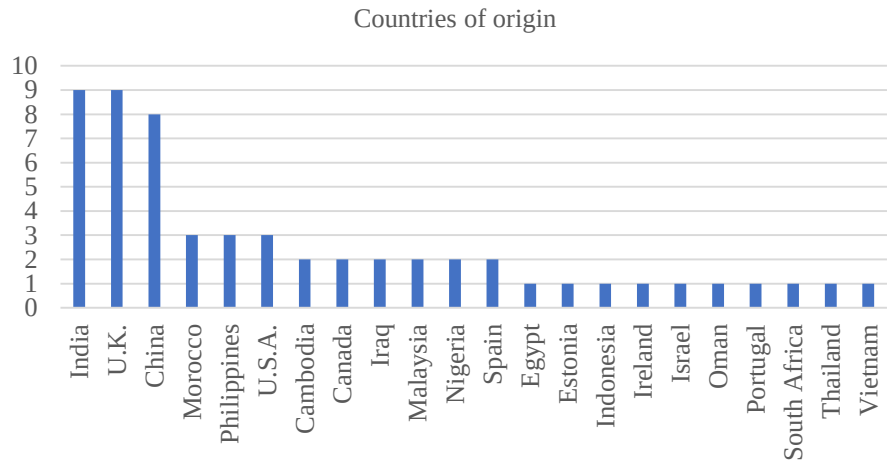


**Fig. 4.** First author's country of origin

**Study context.** The common of the studies, 49 in total, appear to be focused on higher education institutions. The remaining eight articles focused on secondary schools and one article focused on primary school.

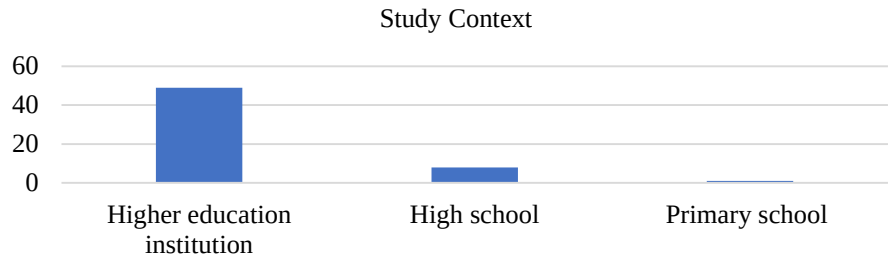


**Fig. 5.** Study context

### 3.2 RQ2: What are the aspects that affect student performance?

This section compiles aspects affecting student performance. Student academic record is the most commonly reported aspect for predicting student performance, accounting for 34% of all compiled aspects, followed by student demographics (26%) and course attributes (11%). These four aspects form 71% of all compiled aspects. A limited number of studies examined other facets, such as student activities, student behavior, instructor attributes, student psychological, and student motivation, which account for the remaining 29%. Figure 6 depicts aspects that influence student performance, as found in the selected articles.

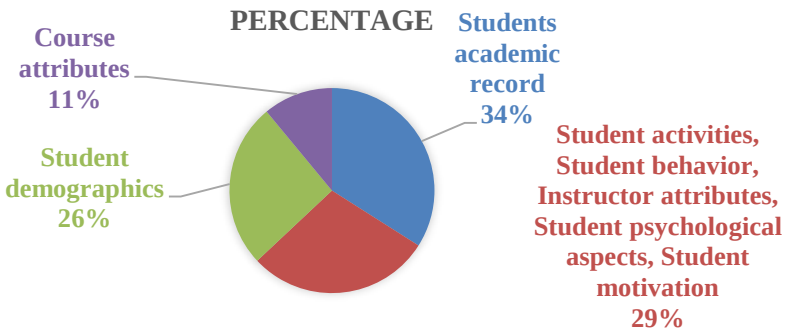


**Fig. 6.** Aspects that affect student performance

### 3.3 RQ3: What are the DM approaches used to predict student performance?

Three DM approaches were discovered in the chosen studies: classification, clustering, and regression. With 48 studies, classification was the most used DM approach. Clustering came in second with 12 studies, and five studies used regression. Two approaches were employed simultaneously in seven (7) studies.

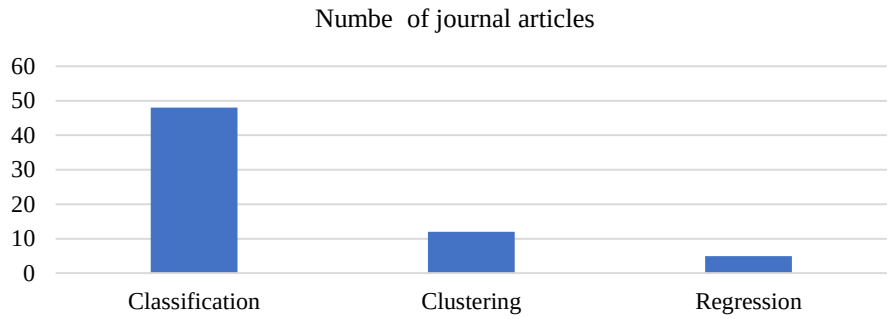


**Fig. 7.** DM approaches

This study discovered 89 DM algorithms based on the 58 chosen articles. These include: LMT, PART, Random Forest (RF), J48, REP Tree, Fine Decision Tree (FDT), OneR, Jrip, Bayes Network (BN), Random Tree, Multi-Layer Perceptron (MLP), ID3, RBF, Decision Tree (DT), Prism, Lasso, K-means kernel, CART, k-Star, Convolutional Neural Networks (CNN), Long Short Term Memory (LSTM), Information gain (IG), Mutual Information (MI), The Proposed FS Method (MICHI), Chi-square (CHI), Linear Regression (LR), Support Vector Machine (SVM), Genetic Algorithm, C4.5, CHAID, Exhaustive CHAID, CRT, QUEST, Linear support vector machines (LSVM), Coarse Decision Tree (CDT), Medium Decision Tree (MDT), Gaussian Naive Bayes (GNB), Kernel Naive Bayes (KNB), Quadratic SVM (QSVM), Cubic SVM (CSVM), Fine Gaussian SVM (FGSVM), Medium Gaussian SVM (MGSVM), Extra Tree, PPP, K-Nearest Neighbors (KNN), Linear support vector machines (L-SVM), Gaussian processes (GP), Neural Network (NN), AdaBoost (ADB), Logistic Regression(LR), mRMR, LLEScorE, SVM-RFE, Relief, ARCO, AVC, MSVM-RFE, ReliefF, MUACD, MAVC, C5.0, Apriori, CER, UBCF, TBR, EduRank, EigenRank, SVD, Dendogram Tree, AIRS2 (Artificial Immune Recognition System v2.0), Deep Thought, Bayesian Additive Regressive Trees (BART), Random Forests (RF), Principal Components Regression (PCR), Multivariate Adaptive Regression Splines (Splines), Tree Ensemble, Resilient backpropagation (Rprop) Multi-Layer Perceptron, Quadratic regression, Fuzzy Analytic Hierarchical Process (AHP), Radial basis function kernel support vector machines (R-SVM), Multiple regression, Swarm Optimization Combined Neural Network, Artificial Neural Network, Adam (Adaptive Moment Estimation), Recurrent Neural Network (RNN), NNge, Back Propagation Neural Network (BP-NN), Deep learning (DL), and Self-Organising Map. Table 5 shows the DM algorithms employed in at least nine studies.

**Table 5.** Main DM algorithms employed

| DM algorithms | Paper ID | Number of articles |
|---|---|---|
| Decision Tree classifiers | P1, P2, P4, P5, P10, P13, P14, P15, P16, P17, P18, P22, P23, P24 P25, P26, P27, P29, P32, P37, P38, P40, P43, P44, P46, P48, P52, P54, P55, P58 | 30 |
| Naïve Bayes classifiers | P1, P6, P7, P10, P17, P19, P22, P24, P35, P36, P37, P54, P58 | 13 |
| Random Forest | P1, P6, P7, P10, P17, P19, P22, P24, P35, P36, P37, P54, P58 | 13 |
| SVM classifiers | P11, P15, P17, P20, P21, P22, P36, P43, P44, P48 | 10 |
| K-Nearest Neighbours | P16, P19, P24, P25, P29, P36, P43, P48, P54 | 9 |

## 4 Discussion

RQ1: What is the research trend of prediction-focus EDM studies?

This section discusses the findings of the review. For RQ1, the two main categories of the research focus are identifying aspects that influence student performance and examining DM algorithms performance. These two categories are often discussed in educational prediction studies in which DM was employed to determine factors that affect student performance and the performance of algorithms used were examined [34]. The findings are consistent with Shin and Shim's research [35]. In addition, DM related to e-learning systems is the third category of the research focus on educational prediction studies as DM can uncover important insight in e-learning environments [36, 37].

This study extends SLRs reported in Shahiri et al. [23] and Saa et al. [22]. Saa et al. [22] analyzed articles published between 2009 and 2018, while Shahiri et al. reviewed articles published between 2002 and 2015. As indicated in Figure 3, there is a surge of publications in this area from 2019 to 2020. Hence, this study is significant to inform the latest trend in prediction-focus EDM studies. Moreover, this study retrieved articles from Lens and Scopus databases. Lens has the largest scholarly record index sourced from Microsoft Academic, Pubmed, and Crossref [38], and it provides comprehensive coverage of all relevant scholarly publications for this SLR study.

RQ2: What are the aspects that affect student performance?

In terms of RQ2, the findings reveal that one of the most influential aspects affecting student performance is student academic records. Thirty-four percent (34%) of the selected studies used student academic records as the attribute to anticipate student performance. In line with Shahiri et al. [23], one-third of their reviewed studies considered CGPA a critical aspect in predicting student performance. Similarly, based on the review of 36 research papers from 2009 to 2018, Saa et al. [23] identified the most common aspects in predicting student performance are students' previous grades and internal assessments. This is further supported by Asif et al. [39] who reported that student performance can be predicted by using academic results without any other characteristics. This aspect is the most used in predicting student performance because it has a measurable value for measuring student performance [40]. In higher education, student

academic records can also be utilized to estimate their graduation [41, 42]. In summary, many prior studies support the finding of this study on this factor.

Another most used aspect to predict student performance is student demographics. Age, ethnicity, gender, housing, family history, and family socioeconomic level are among the student demographics [43]. This finding is in alignment with findings from studies by Hussain et al. [44], Kumar and Radhika [45], Saa et al. [22], and Singh and Pal [46]. Shahiri et al. [23] also discovered that demographic factors are frequently applied to anticipate student performance. Furthermore, according to Hoe et al. [41], there is a strong link between student demographics and academic performance. According to Farooq et al. [47], socioeconomic status is the best indicator of student performance. Apart from school, peer, and student factors, Farooq et al. [47] found that family characteristics are important determinants of students' academic performance. Thus, it can be concluded that student demographics affect student performance.

Eleven percent (11 %) of the studies reviewed used course attributes to predict student performance. This is consistent with Wang and Chung's study [48], which found that course-related variables are predictors of student performance. Other aspects, namely student activities, student behavior, instructor attributes, student psychology, and student motivation, are scarcely reported in ten or fewer studies. According to Saa et al. [22], attributes less reported in the literature are likely to have a minimal influence on student performance prediction. Hence, these aspects have a minor influence on such a prediction.

RQ3: What are the DM approaches used to predict student performance?

Classification is the most applied DM approach in which 83% of the reviewed studies employed this approach, whereas only a few studies employed clustering and regression approaches. This is in line with the study by Mohamad and Tasir [49], which posited that classification is the most used approach in EDM research. As for the DM algorithms, the Decision Tree classifier is the most common algorithm employed in the selected articles. C4.5, J48, and ID3 are examples of decision tree classifiers, which have a tree-like structure with the root node at the top and the leaf nodes at the bottom [50]. Prior studies postulated that Decision Tree classifiers provide a clear way to understand the classification rules [51, 21]. Decision Tree classifiers are frequently used because they are easy to understand and have high predictive accuracy [52].

The second most popular algorithms for predicting student performance are Bayesian Networks and Random Forest. Bayesian Networks are graphical models with nodes and directed edges that are probabilistic in nature [53]. Simple models that explain a certain form of Bayesian network with all attributes being class-conditionally independent are known as Naïve Bayes classifiers [54], which are frequently used in EDM studies [43]. Naïve Bayes classifiers comprise algorithms such as Gaussian, Multinomial, and Bernoulli [22]. The key benefit of employing Naïve Bayes classifiers is that the outcome from the prediction model using Naïve Bayes can be easily translated into human language [43]. Random Forest begins with the conventional Decision Tree. However, this algorithm advances classification to the next level by merging numerous cases [55]. In several studies, Random Forest produces higher accuracy rates when compared with Decision Tree and Naïve Bayes [56, 55]. Next, SVM classifiers rank three in the list of most used algorithms. SVM classifiers are not as regularly used but

it has its advantages. Firstly, SVM has a high degree of generalization and is faster than other methods [57]. SVM is also well-suited for small datasets [16]. Lastly, the K-Nearest Neighbours is identified as the least used algorithm in the selected articles. This algorithm is straightforward in that it maintains all available cases and categorizes new ones using a similarity metric [58]. The benefits of K-Nearest Neighbours include its applicability for both classification and regression predicting issues as well as its robustness in terms of search space where classes do not have to be linearly separable [59].

## 5 Conclusion and future work

This study examined three research questions to gain a better understanding of prediction-focus EDM studies reported in 58 journal articles from 2015 to June 2021. The SLR reveals three categories of the research focus of these 58 studies, which include the identifying aspects influencing student performance (49%), DM algorithms performance (42%), and DM related to e-Learning systems (9%). The selected journal articles are mainly about higher education institutions, and most journal articles were published in 2019 with most authors are from India, United Kingdom, and China. The SLR also reveals student academic records and student demographics as the main aspects used to predict student performance. Classification is the most used DM approach, with Decision Tree classifiers are identified as the most employed algorithm, followed by Bayesian Network and Random Forest. These findings provide useful insights for future researchers and educational practitioners to identify potential and relevant student performance predictors in their respective contexts as well as choices of classifier algorithms for performing such a prediction. A good prediction of student performance also enables early interventions to be implemented to enhance their actual performance.

## 6 Acknowledgements

This work is supported by Kementerian Pengajian Tinggi Malaysia, Fundamental Research Grant Scheme, FRGS/1/2020/SS10/UNIMAS/01/1 and UNIMAS Zamalah Scholarship.

## 7 References

[1] Alfiani, A. P., & Wulandari, F. A. (2015). Mapping student's performance based on data mining approach (a case study). *Agriculture and Agricultural Science Procedia, 3*, 173-177. https://doi.org/10.1016/j.aaspro.2015.01.034

[2] Bari, S., Abdullah, N. A., Abdullah, N., & Yasin, M. H. M. (2016). Early intervention implementation preschool special education students in Malaysia. *International Journal for Innovation Education and Research, 4*(6), 139-155. https://doi.org/10.31686/ijier.vol4.iss7.569

[3] Tan, R. Z., Wang, P. C., Lim, W. H., Ong, S. H. C., & Avnit, K. (2019, December). Early Prediction of Students' Mathematics Performance. *Proceedings of 2018 IEEE International*

*Conference on Teaching, Assessment, and Learning for Engineering (TALE 2018)*, 651-656. https://doi.org/10.1109/TALE.2018.8615289

[4] Aman, F., Rauf, A., Ali, R., Iqbal, F., & Khattak, A. M. (2019, July). A predictive model for predicting students academic performance. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1-4). IEEE. https://doi.org/10.1109/IISA.2019.8900760

[5] Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017, April). Predicting student performance using advanced learning analytics. In *Proceedings of the 26th international conference on world wide web companion* (pp. 415-421). https://doi.org/10.1145/3041021.3054164

[6] McKenzie, K., & Schweitzer, R. (2001). Who Succeeds at University? Factors predicting academic performance in first year Australian university students. *Higher Education Research & Development*, *20*(1), 21–33. https://doi.org/10.1080/07924360120043621

[7] Andujar, P., Bastuji-Garin, S., Botterel, F., Prevel, M., Farcet, J. P., & Claudepierre, P. (2010). Factors affecting students performance on the National Ranking Examination in a French Medical School. *La Presse Medicale*, *39*(6), e134-e140. https://doi.org/10.1016/j.lpm.2010.03.007

[8] Taniguchi, K., Ohashi, K., & Hirakawa, Y. (2013). Analysis of Student's Mathematical Achievement in Grades 3 and 6 in Uganda: Factors Affecting Test Scores and Curriculum Performance. *Procedia-Social and Behavioral Sciences*, *93*, 2058-2062. https://doi.org/10.1016/j.sbspro.2013.10.165

[9] Hand, D. J. (1998). Data mining: statistics and more?. *The American Statistician*, *52*(2), 112-118. https://doi.org/10.1080/00031305.1998.10480549

[10] Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019, March). Student performance prediction and classification using machine learning algorithms. In *Proceedings of the 2019 8th International Conference on Educational and Information Technology* (pp. 7-11). https://doi.org/10.1145/3318396.3318419

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

[12] Adekitan, A. I., & Noma-Osaghae, E. (2019). Data mining approach to predicting the performance of first year student in a university using the admission requirements. *Education and Information Technologies*, *24*(2), 1527-1543. https://doi.org/10.1007/s10639-018-9839-7

[13] Almeda, M. V., Zuech, J., Utz, C., Higgins, G., Reynolds, R., & Baker, R. S. (2018). Comparing the factors that predict completion and grades among for-credit and open/mooc students in online learning. *Online Learning Journal*, *22*(1), 1–18. https://doi.org/10.24059/olj.v22i1.1060

[14] Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*(1), 28-47. https://doi.org/10.1080/21568235.2020.1718520

[15] Özdağoğlu, G., Öztaş, G. Z., & Çağliyangil, M. (2019). An application framework for mining online learning processes through event-logs. *Business Process Management Journal*. https://doi.org/10.1108/BPMJ-10-2017-0279

[16] Zohair, L. M. A. (2019). Prediction of student's performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education*, *16*(1), 27. https://doi.org/10.1186/s41239-019-0160-3

[17] Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning, 19*, 205-220. https://doi.org/10.1007/s10758-014-9223-7

[18] Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3-17.

[19] Bañeres, D., Rodríguez, M. E., Guerrero-Roldán, A. E., & Karadeniz, A. (2020). An Early Warning System to Detect At-Risk Students in Online Higher Education. *Applied Sciences, 10*(13), 4427. https://doi.org/10.3390/app10134427

[20] Atlay, C., Tieben, N., Hillmert, S., & Fauth, B. (2019). Instructional quality and achievement inequality: How effective is teaching in closing the social achievement gap? *Learning and Instruction, 63*, 101211. https://doi.org/10.1016/j.learninstruc.2019.05.008

[21] Maghari, A., & Mousa, H. (2017). School Students' Performance Predication Using Data Mining Classification. *International Journal of Advanced Research in Computer and Communication Engineering, 6*(8), 136-141.

[22] Saa, A. A., Al-Emran, M., & Shaalan, K. (2019). Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. *Technology, Knowledge and Learning, 24*(4), 567-598. Springer Netherlands. https://doi.org/10.1007/s10758-019-09408-7

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

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

[25] Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications, 41*(4), 1432-1462. https://doi.org/10.1016/j.eswa.2013.08.042

[26] Devasia, T., Vinushree, T. P., & Hegde, V. (2016, March). Prediction of students performance using Educational Data Mining. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)* (pp. 91-95). IEEE. https://doi.org/10.1109/SAPIENCE.2016.7684167

[27] Rifat, M. R. I., Al Imran, A., & Badrudduza, A. S. M. (2019). Educational performance analytics of undergraduate business students. *International Journal of Modern Education and Computer Science, 11*(7), 44. https://doi.org/10.5815/ijmecs.2019.07.05

[28] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *Ieee Access, 5*, 15991-16005. https://doi.org/10.1109/ACCESS.2017.2654247

[29] Pati, D., & Lorusso, L. N. (2018). How to write a systematic review of the literature. *HERD, 11*(1), 15-30. https://doi.org/10.1177/1937586717747384

[30] Al-Qaysi, N., Mohamad-Nordin, N., & Al-Emran, M. (2020). A systematic review of social media acceptance from the perspective of educational and information systems theories and models. *Journal of Educational Computing Research, 57*(8), 2085-2109. https://doi.org/10.1177/0735633118817879

[31] Mallett, R., Hagen-Zanker, J., Slater, R., & Duvendack, M. (2012). The benefits and challenges of using systematic reviews in international development research. *Journal of development effectiveness, 4*(3), 445-455. https://doi.org/10.1080/19439342.2012.711342

[32] Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology, 51*(1), 7-15. https://doi.org/10.1016/j.infsof.2008.09.009

[33] Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., ... & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology*, *62*(10), e1-e34. https://doi.org/10.1016/j.jclinepi.2009.06.006

[34] Filiz, E., & Oz, E. (2019). Finding the best algorithms and effective factors in classification of Turkish science student success. *Journal of Baltic Science Education*, *18*(2), 239–253. https://doi.org/10.33225/jbse/19.18.239

[35] Shin, D., & Shim, J. (2021). A systematic review on data mining for mathematics and science education. *International Journal of Science and Mathematics Education*, *19*, 639-659. https://doi.org/10.1007/s10763-020-10085-7

[36] Northcutt, C. G., Ho, A. D., & Chuang, I. L. (2016). Detecting and preventing "multiple-account" cheating in massive open online courses. *Computers & Education*, *100*, 71–80. https://doi.org/10.1016/j.compedu.2016.04.008

[37] Rodrigues, M. W., Isotani, S., & Zárate, L. E. (2018). Educational data mining: A review of evaluation process in the e-learning. *Telematics and Informatics*, *35*(6), 1701–1717. https://doi.org/10.1016/j.tele.2018.04.015

[38] Tay, A. (2018). 7 reasons why you should try Lens.org (updated to version Release 5.16.0 — March 2019). https://aarontay.medium.com/6-reasons-why-you-should-try-lens-org-c40abb09ec6f

[39] Asif, R., Hina, S., & Haque, S. I. (2017). Predicting student academic performance using data mining methods. *International Journal of Computer Science and Network Security (IJCSNS)*, *17*(5), 187–191. http://paper.ijcsns.org/07_book/201705/20170524.pdf

[40] Bin Mat, U., Buniyamin, N., Arsad, P. M., & Kassim, R. (2013, December). An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. In *2013 IEEE 5th conference on engineering education* (ICEED) (pp. 126-130). IEEE. https://doi.org/10.1109/ICEED.2013.6908316

[41] Hoe, A. C. K., Ahmad, M. S., Hooi, T. C., Shanmugam, M., Gunasekaran, S. S., Cob, Z. C., & Ramasamy, A. (2013, November). Analyzing students records to identify patterns of students' performance. In *2013 International Conference on Research and Innovation in Information Systems (ICRIIS)* (pp. 544-547). IEEE. https://doi.org/10.1109/ICRIIS.2013.6716767

[42] Putri, D. Y., Andreswari, R., & Hasibuan, M. A. (2019, August). Analysis of students graduation target based on academic data record using C4. 5 algorithm case study: Information systems students of Telkom University. In *2018 6th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1-6). IEEE. https://doi.org/10.1109/CITSM.2018.8674366

[43] Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The prediction of students' academic performance using classification data mining techniques. *Applied Mathematical Sciences*, *9*(129), 6415-6426. https://doi.org/10.12988/ams.2015.53289

[44] Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447-459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

[45] Kumar, A. D., & Radhika, V. (2014). A survey on predicting student performance. *International Journal of Computer Science and Information Technologies*, *5*(5), 6147-6149.

[46] Singh, R., & Pal, S. (2020). Application of machine learning algorithms to predict students performance. *Int J Adv Sci Technol*, *29*(5), 7249-7261.

[47] Farooq, M. S., Chaudhry, A. H., Shafiq, M., & Berhanu, G. (2011). Factors affecting students' quality of academic performance: A case of secondary school level. *Journal of Quality and Technology Management*, *7*(2), 1-14.
[48] Wang, L., & Chung, S. J. (2021). Sustainable development of college and university education by use of data mining methods. *International Journal of Emerging Technologies in Learning (iJET)*, *15*(5). https://doi.org/10.3991/ijet.v16i05.20303
[49] Mohamad, S. K., & Tasir, Z. (2013). Educational Data Mining: A Review. *Procedia - Social and Behavioral Sciences*, *97*(6), 320-324. https://doi.org/10.1016/j.sbspro.2013.10.240
[50] Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data mining: A prediction for student's performance using classification method. *World Journal of Computer Application and Technology*, *2*(2), 43-47. https://doi.org/10.13189/wjcat.2014.020203
[51] Cheewaprakobkit, P. (2015). Predicting student academic achievement by using the decision tree and neural network techniques. *Human Behavior, Development And Society*, *12*(2), 34-43.
[52] Berhanu, F., & Abera, A. (2015). Students' performance prediction based on their academic record. *International Journal of Computer Applications*, *131*(5), 27–35. https://doi.org/10.5120/ijca2015907348
[53] Uddin, M. F., & Lee, J. (2016). Utilizing Relevant Academic and Personality Features from Big Unstructured Data to Identify Good and Bad Fit Students. *Procedia Computer Science*, *95*, 383-391. https://doi.org/10.1016/j.procs.2016.09.349
[54] Makhtar, M., Nawang, H., & Wan Shamsuddin, S. N. (2017). Analysis on students' performance using naïve bayes classifier. *Journal of Theoretical & Applied Information Technology*, *95*(16).
[55] Kapur, B., Ahluwalia, N., & Sathyaraj, R. (2017). Comparative study on marks prediction using data mining and classification algorithms. *International Journal of Advanced Research in Computer Science*, *8*(3).
[56] Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences*, *10*(11), 3894. https://doi.org/10.3390/app10113894
[57] Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011, April). Prediction of student academic performance by an application of data mining techniques. In *International Conference on Management and Artificial Intelligence IPEDR* (pp. 110-114).
[58] Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, *4*(11). https://doi.org/10.21037/atm.2016.03.37
[59] Khamis, H. S., Cheruiyot, K. W., & Kimani, S. (2014). Application of k-nearest neighbour classification in medical data mining. *International Journal of Information and Communication Technology Research*, *4*(4).
[60] Krishnan, R., Beegum, H., & Sherimon, P.C. (2019). Academic performance based on gender using filter ranker algorithms - An experimental analysis in Sultanate of Oman. *International Journal of Innovative Technology and Exploring Engineering*, *8*(11), 3502-3506. https://doi.org/10.35940/ijitee.K2490.0981119
[61] Sunday, K., Ocheja, P., Hussain, S., Oyelere, S., Samson, B., & Agbo, F. (2020). Analyzing student performance in programming education using classification techniques. *International Journal of Emerging Technologies in Learning (iJET)*, *15*(2), 127-144. https://doi.org/10.3991/ijet.v15i02.11527

[62] Chaudhury, P., & Tripathy, H.K. (2020, September). A novel academic performance estimation model using two stage feature selection. *Indonesian Journal of Electrical Engineering and Computer Science 19*(3), 1610-1619. https://doi.org/10.11591/ijeecs.v19.i3.pp1610-1619

[63] Jamil, J. M., Pauzi, N. F. M., & Nee, I. N. M. S. (2018). An analysis on student academic performance by using decision tree models. *The Journal of Social Sciences Research*, 615-620. https://doi.org/10.32861/jssr.spi6.615.620

[64] Regha, R. S., & Rani, R. U. (2015). A novel clustering based feature selection for classifying student performance. *Indian Journal of Science and Technology*, *8*, 135. https://doi.org/10.17485/ijst/2015/v8iS7/63666

[65] Beemer, J., Spoon, K., Fan, J., Stronach, J., Frazee, J. P., Bohonak, A. J., & Levine, R. A. (2018). Assessing instructional modalities: Individualized treatment effects for personalized learning. *Journal of Statistics Education*, *26*(1), 31-39. https://doi.org/10.1080/10691898.2018.1426400

[66] Thakar, P., & Mehta, A. (2017). A unified model of clustering and classification to improve students' employability prediction. *International Journal of Intelligent Systems and Applications*, *9*(9), 10. https://doi.org/10.5815/ijisa.2017.09.02

[67] Dien, T. T., Luu, S. H., Thanh-Hai, N., & Thai-Nghe, N. (2020). Deep learning with data transformation and factor analysis for student performance prediction. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, *11*(8). https://doi.org/10.14569/IJACSA.2020.0110886

[68] Sokkhey, P., & Okazaki, T. (2020). Developing web-based support systems for predicting poor-performing students using educational data mining techniques. *International Journal of Advanced Computer Science and Applications*, *11(7)*, *23-32.* https://doi.org/10.14569/IJACSA.2020.0110704

[69] Hussain, S., Muhsion, Z. F., Salal, Y. K., Theodorou, P., Kurtoglu, F., & Hazarika, G. C. (2019). Prediction Model on Student Performance based on Internal Assessment using Deep Learning. *International Journal of Emerging Technologies in Learning (iJET)*, *14*(8), 4-22. https://doi.org/10.3991/ijet.v14i08.10001

[70] Khasanah, A., & Harwati, H. (2019). Educational data mining techniques approach to predict students performance. *International Journal of Information and Education Technology*, *9*, 115-118. https://doi.org/10.18178/ijiet.2019.9.2.1184

[71] Bhaskaran, S.S. (2019). Intelligent context driven data mining to analyse student performance in higher educational institutions (HEIs). *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(2), 856-861. https://doi.org/10.35940/ijrte.B1842.078219

[72] Alejandrino, J. C., Delima, A. J. P., & Vilchez, R. N. (2020). IT students selection and admission analysis using naïve bayes and c4. 5 algorithm. *Int. J. Adv. Trends Comput. Sci. Eng*, *9*(1), 759-765. https://doi.org/10.30534/ijatcse/2020/108912020

[73] Evale, D. (2016). Learning management system with prediction model and course-content recommendation module. *Journal of Information Technology Education: Research*, *16*(1), 437-457. https://doi.org/10.28945/3883

[74] Naicker, N., Adeliyi, T., & Wing, J. (2020). Linear Support Vector Machines for Prediction of Student Performance in School-Based Education. *Mathematical Problems in Engineering*, *2020*. https://doi.org/10.1155/2020/4761468

[75] Singh, R., & Pal, S. (2020). Machine learning algorithms and ensemble technique to improve prediction of students performance. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(3), pp. 3970-3976. https://doi.org/10.30534/ijatcse/2020/221932020

[76] Hooshyar, D., Pedaste, M., & Yang, Y. (2020). Mining educational data to predict students' performance through procrastination behavior. *Entropy*, *22*(1), 12. https://doi.org/10.3390/e22010012

[77] Al-Azawei, A., & Al-Masoudy, M. (2020). Predicting Learners' Performance in Virtual Learning Environment (VLE) based on Demographic, Behavioral and Engagement Antecedents. *International Journal of Emerging Technologies in Learning (iJET)*, *15*(9), 60-75. https://doi.org/10.3991/ijet.v15i09.12691

[78] Navamani, J. M. A., & Kannammal, A. (2015). Predicting performance of schools by applying data mining techniques on public examination results. *Research Journal of Applied Sciences, Engineering and Technology*, *9*(4), 262-271. https://doi.org/10.19026/rjaset.9.1403

[79] Aggarwal, D., Mittal, S., & Bali, V. (2019). Prediction Model for Classifying Students Based on Performance using Machine Learning Techniques. *International Journal of Recent Technology and Engineering*, *8*, 496-503. https://doi.org/10.35940/ijrte.B1093.0782S719

[80] Lu, H., & Yuan, J. (2018). Student performance prediction model based on discriminative feature selection. *International Journal of Emerging Technologies in Learning (iJET)*, *13*(10). https://doi.org/10.3991/ijet.v13i10.9451

[81] Wang, X., Yu, X., Guo, L., Liu, F., & Xu, L. (2020). Student performance prediction with short-term sequential campus behaviors. *Information*, *11*(4), 201. https://doi.org/10.3390/info11040201

[82] Polinar, E. L., Delima, A. J. P., & Vilchez, R. N. (2020). Students performance in board examination analysis using naïve bayes and C4. 5 algorithms. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(1), 753-758. https://doi.org/10.30534/ijatcse/2020/107912020

[83] Sokkhey, P., & Okazaki, T. (2020). Study on dominant factor for academic performance prediction using feature selection methods. *International Journal of Advanced Computer Science and Applications*, *11*(8), 492-502. https://doi.org/10.14569/IJACSA.2020.0110862

[84] Yaacob, W. F. W., Nasir, S. A. M., Yaacob, W. F. W., & Sobri, N. M. (2019). Supervised data mining approach for predicting student performance. *Indones. J. Electr. Eng. Comput. Sci*, *16*(3), 1584-1592. https://doi.org/10.11591/ijeecs.v16.i3.pp1584-1592

[85] López-Zambrano, J., Lara, J. A., & Romero, C. (2020). Towards portability of models for predicting students' final performance in university courses starting from Moodle Logs. *Applied Sciences*, *10*(1), 354. https://doi.org/10.3390/app10010354

[86] Segal, A., Gal, K., Shani, G., & Shapira, B. (2019). A difficulty ranking approach to personalization in E-learning. *International Journal of Human-Computer Studies*, *130*, 261-272. https://doi.org/10.1016/j.ijhcs.2019.07.002

[87] Herodotou, C., Rienties, B., Boroowa, A., Zdrahal, Z., & Hlosta, M. (2019). A large-scale implementation of predictive learning analytics in higher education: the teachers' role and perspective. *Educational Technology Research and Development*, *67*(5), 1273-1306. https://doi.org/10.1007/s11423-019-09685-0

[88] Liu, W. (2019). An Improved Back-Propagation Neural Network for the Prediction of College Students' English Performance. *International Journal of Emerging Technologies in Learning (iJET)*, *14*(16). https://doi.org/10.3991/ijet.v14i16.11187

[89] Kovanović, V., Gašević, D., Joksimović, S., Hatala, M., & Adesope, O. (2015). Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. *The Internet and Higher Education*, *27*, 74-89. https://doi.org/10.1016/j.iheduc.2015.06.002

[90] Cheng, F., & Yin, Y. (2019). Application of computer data analysis technology in the development of a physical education examination platform. *International Journal of Emerging Technologies in Learning (iJET)*, *14*(6), 75-86. https://doi.org/10.3991/ijet.v14i06.10158

[91] Hu, C. (2016). Application of e-learning assessment based on AHP-BP algorithm in the cloud computing teaching platform. *International Journal of Emerging Technologies in Learning*, *11*(08), 27-32. https://doi.org/10.3991/ijet.v11i08.6039

[92] Maniktala, M., Cody, C., Barnes, T., & Chi, M. (2020). Avoiding help avoidance: Using interface design changes to promote unsolicited hint usage in an intelligent tutor. *International Journal of Artificial Intelligence in Education*, *30*(4), 637-667. https://doi.org/10.1007/s40593-020-00213-3

[93] Langan, A. M., Harris, W. E., Barrett, N., Hamshire, C., & Wibberley, C. (2018). Benchmarking factor selection and sensitivity: a case study with nursing courses. *Studies in Higher Education*, *43*(9), 1586-1596. https://doi.org/10.1080/03075079.2016.1266613

[94] Howard, E., Meehan, M., & Parnell, A. (2018). Contrasting prediction methods for early warning systems at undergraduate level. *The Internet and Higher Education*, *37*, 66-75. https://doi.org/10.1016/j.iheduc.2018.02.001

[95] Nuankaew, P. (2019). Dropout situation of business computer students, University of Phayao. *International Journal of Emerging Technologies in Learning (iJET)*, *14*(19), 115-131. https://doi.org/10.3991/ijet.v14i19.11177

[96] Sael, N., Hamim, T., & Benabbou, F. (2019). Implementation of the analytic hierarchy process for student profile analysis. *International Journal of Emerging Technologies in Learning (iJET)*, *14*(15). https://doi.org/10.3991/ijet.v14i15.10779

[97] López-Zambrano, J., Lara, J. A., & Romero, C. (2021). Improving the portability of predicting students' performance models by using ontologies. *Journal of Computing in Higher Education*, 1-19. https://doi.org/10.1007/s12528-021-09273-3

[98] Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*, 68-84. https://doi.org/10.1016/j.iheduc.2015.10.002

[99] Rashid, T. A., & Ahmad, H. A. (2016). Lecturer performance system using neural network with Particle Swarm Optimization. *Computer Applications in Engineering Education*, *24*(4), 629-638. https://doi.org/10.1002/cae.21737

[100] Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, *103*, 1-15. https://doi.org/10.1016/j.compedu.2016.09.005

[101] Gamie, E. A., El-Seoud, M., Salama, M. A., & Hussein, W. (2019). Multi-dimensional analysis to predict students' grades in higher education. *International Journal of Emerging Technologies in Learning (iJET)*, *14*(2). https://doi.org/10.3991/ijet.v14i02.9905

[102] Holmes, M., Latham, A., Crockett, K., & O'Shea, J. D. (2017). Near real-time comprehension classification with artificial neural networks: Decoding e-learner non-verbal behavior. *IEEE Transactions on Learning Technologies*, *11*(1), 5-12. https://doi.org/10.1109/TLT.2017.2754497

[103] Crockett, K., Latham, A., & Whitton, N. (2017). On predicting learning styles in conversational intelligent tutoring systems using fuzzy decision trees. *International Journal of Human-Computer Studies*, *97*, 98-115. https://doi.org/10.1016/j.ijhcs.2016.08.005

[104] Parkavi, A., Lakshmi, K., & Srinivasa, K. G. (2017). Predicting effective course conduction strategy using datamining techniques. *Educational Research and Reviews*, *12*(24), 1188-1198. https://doi.org/10.5897/ERR2017.3266

[105] Al-Sudani, S., & Palaniappan, R. (2019). Predicting students' final degree classification using an extended profile. *Education and Information Technologies*, *24*(4), 2357-2369. https://doi.org/10.1007/s10639-019-09873-8

[106] Zhang, Y., & Jiang, W. (2018). Score Prediction Model of MOOCs Learners Based on Neural Network. *International Journal of Emerging Technologies in Learning (iJET)*, *13*(10). https://doi.org/10.3991/ijet.v13i10.9461

[107] Joksimović, S., Gašević, D., Kovanović, V., Riecke, B. E., & Hatala, M. (2015). Social presence in online discussions as a process predictor of academic performance. *Journal of Computer Assisted Learning*, *31*(6), 638-654. https://doi.org/10.1111/jcal.12107

[108] Imran, M., Latif, S., Mehmood, D., & Shah, M. S. (2019). Student Academic Performance Prediction using Supervised Learning Techniques. *International Journal of Emerging Technologies in Learning (iJET)*, *14*(14). https://doi.org/10.3991/ijet.v14i14.10310

[109] Yang, F., & Li, F. W. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education*, *123*, 97-108. https://doi.org/10.1016/j.compedu.2018.04.006

[110] Wakelam, E., Jefferies, A., Davey, N., & Sun, Y. (2020). The potential for student performance prediction in small cohorts with minimal available attributes. *British Journal of Educational Technology*, *51*(2), 347-370. https://doi.org/10.1111/bjet.12836

[111] El Mabrouk, M., Gaou, S., & Rtili, M. K. (2017). Towards an Intelligent Hybrid Recommendation System for E-Learning Platforms Using Data Mining. *International Journal of Emerging Technologies in Learning (iJET)*, *12*(6). https://doi.org/10.3991/ijet.v12i06.6610

[112] Belarbi, N., Chafiq, N., Talbi, M., Namir, A., & Benlahmar, E. (2019). User Profiling in a SPOC: A method based on User Video Clickstream Analysis. *International Journal of Emerging Technologies in Learning (iJET)*, *14*(1). https://doi.org/10.3991/ijet.v14i01.9091

[113] Ognjanovic, I., Gasevic, D., & Dawson, S. (2016). Using institutional data to predict student course selections in higher education. *The Internet and Higher Education*, *29*, 49-62. https://doi.org/10.1016/j.iheduc.2015.12.002

[114] Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education*, *131*, 22-32. https://doi.org/10.1016/j.compedu.2018.12.006

## 8 Authors

**Muhammad Haziq bin Roslan** is currently a postgraduate student at the Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak. His areas of research interest include learning sciences, MOOC, and educational data mining.

**Chwen Jen Chen** is a professor attached to the Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak. Her scholarly work focuses on fundamental and applied aspects of learning design and learning technology. Her recent research foci also include educational data mining and students-as-partners in online learning environments.

Article submitted 2021-10-19. Resubmitted 2021-12-16. Final acceptance 2021-12-20. Final version published as submitted by the authors.

# 9 Appendix

**Table 6.** Selected papers

| Paper ID | Title | Journal | Author | Country | Year | Context | Ref. |
|---|---|---|---|---|---|---|---|
| P1 | Academic Performance Based on Gender using Filter Ranker Algorithms - An Experimental Analysis in Sultanate of Oman | International Journal of Innovative Technology and Exploring Engineering | Krishnan, R., Beegum, H., Sherimon, P.C. | Oman | 2019 | Higher education institution | [60] |
| P2 | Analyzing Student Performance in Programming Education Using Classification Techniques | International Journal of Emerging Technologies in Learning | Sunday, K., Ocheja, P., Hussain, S., (...), Balogun, O.S., Agbo, F.J. | Nigeria | 2020 | Higher education institution | [61] |
| P3 | A novel academic performance estimation model using two stages feature selection | Indonesian Journal of Electrical Engineering and Computer Science | Chaudhury, P., Tripathy, H.K. | India | 2020 | Higher education institution | [62] |
| P4 | An Analysis on Student Academic Performance by Using Decision Tree Models | Journal of Social Sciences Research | Jamil, J.M., Pauzi, N.F.M., Nee, I.N.M.S. | Malaysia | 2018 | Higher education institution | [63] |
| P5 | A Novel Clustering based Feature Selection for Classifying Student Performance | Indian Journal of Science and Technology | Regha, R. S., Uma Rani, R. | India | 2015 | Higher education institution | [64] |
| P6 | Assessing Instructional Modalities: Individualized Treatment Effects for Personalized Learning | Journal of Statistics Education | Beemer, J., Spoon, K., Fan, J., (...), Bohonak, A.J., Levine, R.A. | United States of America | 2018 | Higher education institution | [65] |
| P7 | A unified model of clustering and classification to improve students' employability prediction | International Journal of Intelligent Systems and Applications | Thakar, P., Mehta, A., Manisha | India | 2017 | Higher education institution | [66] |
| P8 | Deep Learning with Data Transformation and Factor Analysis for Student Performance Prediction | International Journal of Advanced Computer Science and Applications | Dien, T.T., Luu, S.H., Thanh-Hai, N., Thai-Nghe, N. | Vietnam | 2020 | Higher education institution | [67] |
| P9 | Developing Web-based Support Systems for Predicting Poor-performing Students using Educational Data Mining Techniques | International Journal of Advanced Computer Science and Applications | Sokkhey, P., Okazaki, T. | Cambodia | 2020 | High school | [68] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P10 | Educational Data Mining and Analysis of Students' Academic Performance Using WEKA | Indonesian Journal of Electrical Engineering and Computer Science | Hussain, S., Dahan, N.A., Ba-Alwib, F.M., Ribata, N. | India | 2018 | Higher education institution | [69] |
| P11 | Educational Data Mining Techniques Approach to Predict Student's Performance | International Journal of Information and Education Technology | Khasanah, A.U., Harwati, H. | Indonesia | 2019 | Higher education institution | [70] |
| P12 | Intelligent Context Driven Data Mining To Analyse Student Performance in Higher education institutional Institutions (HEIs) | International Journal of Recent Technology and Engineering (IJRTE) | Bhaskaran, S.S. | Bahrain | 2019 | Higher education institution | [71] |
| P13 | IT Students Selection and Admission Analysis using Naïve Bayes and C4.5 Algorithm | International Journal of Advanced Trends in Computer Science and Engineering | Alejandrino, J.C., Delima, A.J.P., Vilchez, R.N. | Philippines | 2020 | High school | [72] |
| P14 | Learning Management System With Prediction Model And Course-Content Recommendation Module | Journal of Information Technology Education: Research | Evale, D.S. | Philippines | 2017 | Higher education institution | [73] |
| P15 | Linear Support Vector Machines for Prediction of Student Performance in School-Based Education | Mathematical Problems in Engineering | Naicker, N., Adeliyi, T., Wing, J. | South Africa | 2020 | High school | [74] |
| P16 | Machine Learning Algorithms and Ensemble Technique to Improve Prediction of Students Performance | International Journal of Advanced Trends in Computer Science and Engineering | Singh, R., Pal, S. | India | 2020 | Higher education institution | [75] |
| P17 | Mining Educational Data to Predict Students' Performance through Procrastination Behavior | Entropy | Hooshyar, D., Pedaste, M., Yang, Y. | Estonia | 2020 | Higher education institution | [76] |
| P18 | Predicting Learners' Performance in Virtual Learning Environment (VLE) based on Demographic, Behavioral and Engagement Antecedents | International Journal of Emerging Technologies in Learning (iJET) | Al-Azawei, A., Al-Masoudy, M.A.A. | Iraq | 2020 | Higher education institution | [77] |
| P19 | Predicting Performance of Schools by Applying Data Mining Techniques on Public Examination Results | Research Journal of Applied Sciences, Engineering and Technology | Navamani, J.M.A., Kannammal, A. | India | 2015 | High school | [78] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P20 | Prediction Model for Classifying Students Based on Performance using Machine Learning Techniques | International Journal of Recent Technology and Engineering | Aggarwal, D., Mittal, S., Bali, V. | India | 2019 | Higher education institution | [79] |
| P21 | Student Performance Prediction Model Based on Discriminative Feature Selection | International Journal of Emerging Technologies in Learning | Lu, H., Yuan, J. | China | 2018 | High school | [80] |
| P22 | Student Performance Prediction with Short-Term Sequential Campus Behaviors | Information (Switzerland) | Wang, X., Yu, X., Guo, L., Liu, F., Xu, L. | China | 2020 | Higher education institution | [81] |
| P23 | Students Performance in Board Examination Analysis using Naïve Bayes and C4.5 Algorithms | International Journal of Advanced Trends in Computer Science and Engineering | Polinar, E.L., Delima, A.J.P., Vilchez, R.N. | Philippines | 2020 | Higher education institution | [82] |
| P24 | Study on Dominant Factor for Academic Performance Prediction using Feature Selection Methods | International Journal of Advanced Computer Science and Applications | Sokkhey, P., Okazaki, T. | Cambodia | 2020 | High school | [83] |
| P25 | Supervised data mining approach for predicting student performance | Indonesian Journal of Electrical Engineering and Computer Science | Yaacob, W.F.W., Nasir, S.A.M., Yaacob, W.F.W., Sobri, N.M. | Malaysia | 2019 | Higher education institution | [84] |
| P26 | Sustainable Development of College and University Education by Use of Data Mining Methods | International Journal of Emerging Technologies in Learning | Wang, L., Chung, S.-J. | China | 2021 | Higher education institution | [48] |
| P27 | Towards Portability of Models for Predicting Students' Final Performance in University Courses Starting from Moodle Logs | Applied Sciences (Switzerland) | López-Zambrano, J., Lara, J., Romero, C. | Spain | 2020 | Higher education institution | [85] |
| P28 | A Difficulty Ranking Approach to Personalization in E-learning | International Journal of Human-Computer Studies | Segal, A., Gal, K., Shani, G., Shapira, B. | Israel | 2019 | Primary school | [86] |
| P29 | A large‐scale implementation of predictive learning analytics in higher education: the teachers' role and perspective | Educational Technology Research and Development | Herodotou, C., Rienties, B., Boroowa, A., Zdrahal, Z., & Hlosta, M. | United Kingdom | 2019 | Higher education institution | [87] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P30 | An Improved Back-propagation Neural network for the Prediction of College Students' English Performanc | International Journal of Emerging Technologies in Learning | Liu, W. | China | 2019 | Higher education institution | [88] |
| P31 | Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions | The Internet and Higher Education | Kovanović, V., Gašević, D., Joksimović, S., Hatala, M., & Adesope, O. | Canada | 2015 | Higher education institution | [89] |
| P32 | Application of Computer Data Analysis Technology in the Development of a Physical Education Examination Platform | International Journal of Emerging Technologies in Learning | Cheng, F., Yin, Y. | China | 2019 | Higher education institution | [90] |
| P33 | Application of E-Learning Assessment Based on AHP-BP Algorithm in the Cloud Computing Teaching Platform | International Journal of Emerging Technologies in Learning | Hu, C. | China | 2016 | Higher education institution | [91] |
| P34 | Avoiding Help Avoidance: Using Interface Design Changes to Promote Unsolicited Hint Usage in an Intelligent Tutor | International Journal of Artificial Intelligence in Education | Maniktala, M., Cody, C., Barnes, T., Chi, M. | United States of America | 2020 | Higher education institution | [92] |
| P35 | Benchmarking factor selection and sensitivity: a case study with nursing courses | Studies in Higher Education | Langan, A. M., Harris, W. E., Barrett, N., Hamshire, C., & Wibberley, C. | United Kingdom | 2019 | Higher education institution | [93] |
| P36 | Contrasting prediction methods for early warning systems at undergraduate level | The Internet and Higher Education | Howard, E., Meehan, M., Parnell, A. | Ireland | 2018 | Higher education institution | [94] |
| P37 | Data mining approach to predicting the performance of first year student in a university using the admission requirements | Education and Information Technologies | Adekitan, A. I., Noma-Osaghae, E. | Nigeria | 2018 | Higher education institution | [12] |
| P38 | Dropout Situation of Business Computer Students, University of Phayao | International Journal of Emerging Technologies in Learning | Nuankaew, P. | Thailand | 2019 | Higher education institution | [95] |
| P39 | Implementation of the Analytic Hierarchy Process for Student Profile Analysis | International Journal of Emerging Technologies in Learning | Sael, N., Hamim, T., Benabbou, F. | Morocco | 2019 | Higher education institution | [96] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P40 | Improving the portability of predicting students' performance models by using ontologies | Journal of Computing in Higher Education | López-Zambrano, J., Lara, J. A., Romero, C. | Spain | 2021 | Higher education institution | [97] |
| P41 | Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success | The Internet and Higher Education | Gašević, D., Dawson, S., Rogers, T., Gasevic, D. | United Kingdom | 2019 | Higher education institution | [98] |
| P42 | Lecturer Performance System Using Neural Network with Particle Swarm Optimization | Computer Applications in Engineering Education | Rashid, T. A., Ahmad, H. A. | Iraq | 2016 | Higher education institution | [99] |
| P43 | Models for early prediction of at-risk students in a course using standards-based grading | Computers & Education | Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. | United States of America | 2016 | Higher education institution | [100] |
| P44 | Multi-Dimensional Analysis to Predict Students' Grades in Higher Education | International Journal of Emerging Technologies in Learning | Gamie, E. A., El-Seoud, M., Salama, M. A., Hussein, W. | Egypt | 2019 | Higher education institution | [101] |
| P45 | Near real-time comprehension classification with artificial neural networks: decoding e-Learner non-verbal behavior | IEEE Transactions on Learning Technologies | Holmes, M., Latham, A., Crockett, K., & O'Shea, J. D. | United Kingdom | 2017 | Higher education institution | [102] |
| P46 | On Predicting Learning Styles in Conversational Intelligent Tutoring Systems using Fuzzy Decision Trees | International Journal of Human-Computer Studies | Crockett, K., Latham, A., Whitton, N. | United Kingdom | 2017 | Higher education institution | [103] |
| P47 | Predicting effective course conduction strategy using data mining techniques | Educational Research and Reviews | Parkavi, A., Lakshmi, K., Srinivasa, K. G. | India | 2017 | Higher education institution | [104] |
| P48 | Predicting students' final degree classification using an extended profile | Education and Information Technologies | Al-Sudani, S., Palaniappan, R. | United Kingdom | 2019 | Higher education institution | [105] |
| P49 | Prediction Model on Student Performance based on Internal Assessment using Deep Learning | International Journal of Emerging Technologies in Learning | Hussain, S., Muhsion, Z. F., Salal, Y. K., Theodorou, P., Kurtoglu, F., Hazarika, G. C. | India | 2019 | Higher education institution | [69] |
| P50 | Score Prediction Model of MOOCs Learners Based on Neural Network | International Journal of Emerging Technologies in Learning | Zhang, Y., Jiang, W. | China | 2018 | Higher education institution | [107] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P51 | Social presence in online discussions as a process predictor of academic performance | Journal of Computer Assisted Learning | Joksimović, S., Gašević, D., Kovanović, V., Riecke, B. E., & Hatala, M. | United Kingdom | 2015 | Higher education institution | [107] |
| P52 | Student Academic Performance Prediction using Supervised Learning Techniques | International Journal of Emerging Technologies in Learning | Imran, M., Latif, S., Mehmood, D., & Shah, M. S. | Portugal | 2019 | High school | [108] |
| P53 | Study on Student Performance Estimation, Student Progress Analysis, and Student Potential Prediction based on Data Mining | Computers & Education | Yang, F., & Li, F. W. | China | 2018 | High school | [109] |
| P54 | The potential for student performance prediction in small cohorts with minimal available attributes | British Journal of Educational Technology | Wakelam, E., Jefferies, A., Davey, N., Sun, Y. | United Kingdom | 2020 | Higher education institution | [110] |
| P55 | Towards an Intelligent Hybrid Recommendation System for E-Learning Platforms Using Data Mining | International Journal of Emerging Technologies in Learning | El Mabrouk, M., Gaou, S., Rtili, M. K. | Morocco | 2017 | Higher education institution | [111] |
| P56 | User Profiling in a SPOC: A method based on User Video Clickstream Analysis | International Journal of Emerging Technologies in Learning | Belarbi, N., Chafiq, N., Talbi, M., Namir, A., Benlahmar, E. | Morocco | 2019 | Higher education institution | [112] |
| P57 | Using institutional data to predict student course selections in higher education | The Internet and Higher Education | Ognjanovic, I., Gasevic, D., Dawson, S. | Canada | 2016 | Higher education institution | [113] |
| P58 | Utilizing Early Engagement and Machine Learning to Predict Student Outcomes | Computers & Education | Gray, C. C., Perkins, D. | United Kingdom | 2019 | Higher education institution | [114] |

**Table 7.** Data extraction

| Paper ID | Factors examined | DM algorithms |
|---|---|---|
| P1 | Gender, Students' Income, Students' Transportation, Mode of study, Accommodation Type, Job status, GPA, Result of the student, Size of the class | J48, LMT, Random Forest, Random Tree, REP Tree, OneR, PART, Jrip, Bayes Network, Multi-Layer Perceptron |
| P2 | First semester grade, Class test score, Second semester grade, Assignment completed, Class lab work, Class attendance | ID3, J48 |
| P3 | Gender, Background (urban, rural), Mother's qualification, Mother's occupation, Father's qualification, Father's occupation, Parents staying together or apart, Joined engineering willingly or under parental pressure, Quality of family relations, Most engineering subjects | RBF |

| | | |
|---|---|---|
| | are interesting, Most engineering subjects are complex, Participation in extracurricular activities, Time spent in sports or extra-curricular activities, Study duration weekly, Time spent with friends weekly, Sleeping time duration daily, Financial distress, Health impairment, Type of internet connectivity on smartphone, Frequency of internet usage daily, Frequency of internet access during class, Does internet feel addictive, Absent percentage, Previous failures, 10 class division, 12 class division, Internal grade1, Internal grade 2, Previous CGPA1, Previous CGPA 2 | |
| P4 | Age, Ethnics, Gender, Parents Income, Entry Level, Program, CGPA | Decision Tree |
| P5 | Students' name, age, gender, family belong to nuclear family or joint family, family occupation & educational, qualification of family members, economic factors, location feature, government, self-financed, personal factors, course and nature of college, college factors, social factors and spending time in television, mobile, computer, Academic factors | Prism, J48 |
| P6 | GPA, Experience with online courses, Age | Random Forests, Lasso |
| P7 | Gender, Age, State, Graduation Univ, School state 10, School state 12, Graduation state, $10^{th}$, $12^{th}$, Ug/Pg sem1, Ug/Pg sem2, Graduation per, Stream12, Mother occupation, Father occupation, Mother qualification, Father qualification, Type of family, Number of siblings, Type of school 10, Type of school 12, Regular or distant, Sales result oriented, Sales multitask, Sales inquisitive, Sales flexible, Sales charismatic, Sales people, Cognitive skills articulate, Cognitive skills proficient, Cognitive skills logical, HrScore, HrPer, Automata score, Automata per, Computer prog, Score CS score, Excel score, Computer Prog Per, Fascore, faper, English Score, English Per, Vocabulary, Grammar, Comprehension, Communication, Quant Score, QA per, Logical Score, LA per, Inductive Reasoning, Deductive Reasoning, Abductive Reasoning, Basic Mathematics, Engineering Mathematics, Applied Mathematics, Personality Score, Personality Per, Agreeableness, Extraversion, Neuroticism, Conscientiousness | K-means kernel, Simple CART, k-Star, Random Tree, Random Forest |
| P8 | CGPA, CGPA-Pre-Semester, Credits earned, English Mark_L1, English Mark_L2, English Mark_L3, Entrance Mark s1, Entrance Mark s2, Entrance Mark s3, GPA Semester, Students' intake, Faculty, Field of Study, Number of credits, Gender | Long Short Term Memory (LSTM), Convolutional Neural Networks (CNN) |
| P9 | Parents' educational levels, Parents' occupational status, Parents' socioeconomic levels, Parents' involvement, Parenting styles, Domestic environment, Self-disciplines, Students' interest and motivation, Students' anxiety toward their classes and exams, Students' possession materials, Class environment, Curriculum, Teaching methods and practices, Teachers' attribute & characteristics, Academic resource | Chi-square (CHI), Information gain (IG), Mutual Information (MI), The Proposed FS Method (MICHI) |
| P10 | Gender, Caste, Class X Percentage, Class XII Percentage, Internal assessment percentage, Marital status, Lived in town or village, Admission category, Family | J48, PART, Random Forest, Bayes Network |

| | | |
|---|---|---|
| | monthly income, Family size, Father qualification, Mother qualification, Father occupation, Mother Occupation, Number of friends, Study hours, Student school attended at Class X level, Medium, home to college Travel time, Class attendance percentage | |
| P11 | Senior high school grade, GPA in first semester, Final GPA, Attendance in first semester | K-means, Linear Regression, Support Vector Machine |
| P12 | GPA, Time to degree, course code, Semester, course difficulty, Course complexity, Student potential, Course weight, Course type, Course level, Class size, Student prior learning details | Genetic Algorithm |
| P13 | Sex, Status, address, Age, Citizenship, Economic status, Strand, Course first choice, GPA, Verbal, Non-verbal, Stanine, Math, Science, English, Standing | C4.5, Naive Bayes |
| P14 | Gender, Age, Course, section, Schedule, Grade in Programming 1, Grade in Programming 2, Grade in Programming 3 | CHAID, Exhaustive CHAID, CRT, QUEST, J48, Bayes Network, Naive Bayes, JRip |
| P15 | Gender, Race/ethnicity, Parental level of education, Access to lunch, Test preparation, Mathematics score, Reading score, Writing score | Linear support vector machines (LSVM), Fine Decision Tree (FDT), Coarse Decision Tree (CDT), Medium Decision Tree (MDT), logistic regression (LR), Gaussian Naive Bayes (GNB), Kernel Naive Bayes (KNB), quadratic SVM (QSVM), cubic SVM (CSVM), fine Gaussian SVM (FGSVM), medium Gaussian SVM (MGSVM) |
| P16 | Sex of students, Students category, Discussion at home, Own computer /laptop, Laptop shared with family, Study desk at home, Own mobile phone, Own Gaming system, Heating/Cooling systems at, Absent from school, How often use computer at school, Access textbooks, Completed assignments, Collaborate with classmates, Communicate with teacher, Students grade in Senior Secondary Education, Fathers qualification, Mother's qualification, Father's occupation, Mother's occupation, Performance in B.C.A | Decision Tree, Naïve Bayes, K-Nearest Neighbors, Extra Tree |
| P17 | Open date of an assignment (OpenD), the date of first view of the assignment (FirstviewD), the date of assignment submission (SubmissionD), the due date of the assignment (Deadline) | PPP, Linear support vector machines (L-SVM), radial basis function kernel support vector machines (R-SVM), Gaussian processes (GP), Decision Tree, Random Forest, Neural Network, AdaBoost (ADB), Naive Bayes |
| P18 | Gender, region, highest education, deprivation band, age band, disability, the number of previous attempts | Decision Tree, Linear Regression, M5P Regression |
| P19 | Public examinations results | Naïve Bayes, Random Forest, K-Nearest Neighbors |
| P20 | Gender, Caste, Matric Percentage, XII Percentage, Reappear/back paper, Marital status, Living status, Admission category, Family income, Family size, Father's qualification, Mother's qualification, Father's occupation, Mother's occupation, Number of friends, Study hours, Type of school attended, Medium, Travel time between college and home | Naive Bayes, Logistic Regression, Support Vector Machine, Multi-Layer Perceptron, J48, Random Forest |

| | | |
|---|---|---|
| P21 | School, sex, age, address, Family size, Parent's cohabitation status, Mother's education, Father's education, Mother's job, Father's job, Reason to choose this school, Student's guardian, Home to school travel time, Family educational support, Quality of family relationships, Number of past failures, school educational support, Off-campus tutorial, extra-curricular activities, Attended nursery school, Wants to take higher education, Number of school absences, The first monthly exam grade, Midterm exam grade, The second monthly exam grade, Weekly study time, Internet access at home, With a romantic relationship, Free time after school, Going out with friends, Workday alcohol consumption, Weekend alcohol consumption, Current health status | mRMR, LLEScorE, SVM-RFE, Relief, ARCO, AVC, MSVM-RFE, ReliefF, MUACD, MAVC, Support Vector Machine |
| P22 | Books lending time, books' name, Books' ISBN, Category of consumption, Position of consumption, Way of consumption, Time of consumption, Amount of consumption, Balance of consumption, Students' id, Time of entering/leaving dormitory, Direction of entering/leaving dormitory, Number of library gate, Time of entering/leaving library, Number of faculty, Grades ranking | Support Vector Machine, Logistic Regression, Bayes Network, Decision Tree, Random Forest |
| P23 | Student stanine during the admission test, Undergraduate Student General Weighted Average, Student scholarship while in college, Student honors received during graduation, A student who enrolled in LET Review Centers, A student who passed/failed the LET Exam | Naive Bayes, C4.5 |
| P24 | Parents' educational levels, Parents' occupational status, Parents' socioeconomic levels, Parents' involvement, Parenting styles, Domestic environment, Self-disciplines, Students' interest and motivation, Students' anxiety toward their classes and exams, Students' possession materials, School and class environment, Curriculum, Teaching methods and practices, Teachers' attribute & characteristics, Academic resource, Student's performance level based on their mark or score | K-Nearest Neighbor, C5.0, Random Forest |
| P25 | Gender, Final CGPA, All the courses enrolled by the students, Courses' grades | K-Nearest Neighbor, Naïve Bayes, Decision Tree, Logistic Regression |
| P26 | Courses' score, Lecturers' name, Courses' information | Apriori, Decision Tree |
| P27 | Events in the log file, Final mark in the courses | J48 |
| P28 | Grades, Number of retries, Time spent solving questions | CER, UBCF, TBR, EduRank, EigenRank SVD |
| P29 | Gender, Age, Disability, Ethnicity, Education level, Index of Multiple Deprivation (IMD band), Students' previous experience of studying at a university, Best previous course score achieved, Sum of previous credits achieved | Naïve Bayes, CART, k-Nearest Neighbors |
| P30 | NCEE Score, Gender, Age, Learning attitude | Neural Network |
| P31 | Technology-use profiles, Cognitive presence | Dendogram Tree |
| P32 | Exam type, Exam item, Score management, Score analysis, System administration | Decision tree |
| P33 | Learning attitude, Capability of active participation, The use of E-learning platform, Capability of cooperative learning | Neural Network |

| | | |
|---|---|---|
| P34 | Impact of completion rate, System errors on the group | Deep Thought |
| P35 | Age, Gender, Ethnicity, Registered 'disability', Home address | Random Forests |
| P36 | Gender, Elective, Option or core, Repeating course, Students' year of study, Students' program | Bayesian Additive Regressive Trees (BART), Random Forests (RF), Principal Components Regression (PCR), Multivariate Adaptive Regression Splines (Splines), K-Nearest Neighbours (KNN), Neural Networks (NN), Support Vector Machine (SVM) |
| P37 | CGPA, Class of degree | Random Forest, Tree Ensemble, Decision Tree, Naive Bayes, Logistic Regression, Resilient backpropagation (Rprop) Multi-Layer Perceptron, Linear regression, Quadratic regression |
| P38 | Students, Attendance class, Join the activities | Decision Tree |
| P39 | Academic results, Basic information, Financial situation, Intellectual level of parents, Information on family stability | Fuzzy Analytic Hierarchical Process (AHP) |
| P40 | Subject or name of the course, Identification code, Name of the degree, Year in the degree/curriculum, number of students, Level of Moodle Usage | J48 |
| P41 | Percent mark, Academic status, Age, Gender, International student, Language spoken at home, Home remoteness, Previous enrollment in the same course | Multiple regression |
| P42 | Student Feedback, Lecturers Portfolio, Continuous Academic Development (CAD) | Swarm Optimization Combined Neural Network |
| P43 | Grades for attendance, Quizzes, Weekly homework and participation, Project milestones, Mathematical modeling activity tasks, Exams | Logistic Regression, Support Vector Machine, Decision Tree, Multi-Layer Perceptron, Naive Bayes, K-Nearest Neighbour |
| P44 | Student grade, Number of course logins, School leaving grade, Module type, Attendance | Neural Networks, Decision Tree, Support Vector Machine, Bayesian network |
| P45 | Detecting learner comprehension of on-screen information during e-learning activities | Artificial Neural Network |
| P46 | Behaviour, Test scores | Decision Tree |
| P47 | Course conduction strategy | Linear regression |
| P48 | Academic, Demographic, Psychological, Economic | Neural Network, K-Nearest Neighbour, Decision Tree, Support Vector Machine |
| P49 | Exam, Subject, Internal assessment marks, CGPA, Result | Adam (Adaptive Moment Estimation), Deep learning (DL), Recurrent Neural Network (RNN), AIRS2 (Artificial Immune Recognition System v2.0), Adaboost |
| P50 | Learner learning behaviour, Course interaction times, Number of interactive days in the course, Number of course chapters, Number of posts in the forum, The length of course | Neural Network |
| P51 | Social presence | Multiple regression |
| P52 | Educational background, Social, Demographics, Family, Socioeconomic status | Decision Tree, NNge, MLP |

| | | |
|---|---|---|
| P53 | Learning Mode, Perception, Input, Organization, Processing, Understanding | Back Propagation Neural Network (BP-NN) |
| P54 | Age, Gender, Region, Residence, Guardian info, Cleared certificates, Scholarships, Results, Recent assignment results, Quizzes, Final exam, CGPA, Attendance, Interaction with social media websites, Games partitions, Sports, Hobbies, Behavior, Absence, Remarks | Decision Tree (DT), K-Nearest Neighbours (KNN), Random Forest (RF) |
| P55 | Field, Subject, Publishing house, Book | Decision Tree |
| P56 | Video clickstream behaviour | Bayesian method, K-means algorithm |
| P57 | Prior course enrolments, GPA, Gender, Country of origin, Potential career objectives, Course scheduling, Instructor demographics, Course, Teacher evaluations | AHP based algorithm |
| P58 | Attendance, Academic standing, School, Program, Year | Random Tree, Random Forest, Naive Bayes, Multi-layer Perceptron, Self-Organising Map, C4.5 Tree |

**Table 8.** Research focus of the selected articles

| Research Topics | Paper ID |
|---|---|
| Identification of attributes influencing students' performance | P1, P2, P3, P4, P7, P10, P13, P14, P15, P16, P17, P19, P21, P22, P23, P24, P25, P26, P27, P29, P30, P31, P32, P35, P36, P37, P38, P39, P41, P46, P47, P48, P50, P51, P53, P54, P56, P57, P58 |
| Data mining performance | P5, P8, P9, P10, P11, P12, P14, P15, P16, P19, P20, P21, P23, P24, P25, P26, P32, P33, P34, P36, P37, P39, P40, P42, P43, P44, P45, P46, P47, P48, P49, P52, P53 |
| E-learning systems | P6, P18, P28, P33, P45, P50, P55 |

**Table 9.** Categories of factors that affect the students' performance and their description

| Attributes affecting the students' academic success | Description | Paper ID |
|---|---|---|
| Students' academic Record | The records that related to academic performance e.g., study duration, student's exam result, assessment mark, GPA, CGPA, and class attendance. | P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13, P14, P15, P16, P18, P19, P20, P21, P22, P23, P24, P25, P27, P28, P29, P30, P37, P38, P39, P41, P42, P43, P44, P46, P48, P49, P52, P54, P57, P58 |
| Students' demographics | The demographic factors include age, gender, geographical affiliation, ethnicity, nationality, marital status, socioeconomic status (SES), parental education, language, financial, and religious affiliations. | P1, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13, P14, P15, P16, P18, P20, P21, P22, P24, P25, P28, P29, P30, P35, P36, P39, P41, P48, P52, P54, P57 |
| Students' activities | Information that related to the students' activities such as extracurricular activities, e-learning activities, internet usage and time spent with friends. | P3, P5, P16, P21, P31, P33, P45, P51, P52, P54 |
| Students' behaviour | The way that the students behave e.g., procrastination and self-discipline. | P17, P18, P22, P24, P30, P33, P34, P50, P56 |
| Students' motivation | Factors that make the students be motivated e.g., parents' involvement and students' interest. | P3, P24 |

| | | |
|---|---|---|
| Courses' attributes | Characteristics that related to the course e.g., credit hours, field of study and course difficulty. | P4, P8, P26, P32, P36, P40, P44, P47, P49, P50, P55, P57, P58 |
| Instructors' attributes | Characteristics that related to the instructor e.g., teaching methods and practices. | P9, P24, P26, P27, P29, P42, P57 |
| Students' psychological aspects | Psychological attributes of the students e.g., cognitive abilities and personalities. | P7, P9, P31, P39, P48, P53, P54 |

**Table 10.** Categories of DM approaches

| **Categories of DM approaches** | **Paper ID** |
|---|---|
| Classification | P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13, P14, P15, P16, P17, P18, P19, P20, P21, P22, P23, P24, P25, P26, P27, P29, P30, P32, P33, P35, P37, P38, P39, P40, P42, P43, P44, P45, P46, P48, P49, P52, P53, P54, P55, P56, |
| Clustering | P7, P11, P17, P28, P31, P34, P36, P49, P50, P56, P57, P58 |
| Regression | P18, P37, P41, P47, P51 |