ISSN: 2633-4828 Vol. 5 No.4, December, 2023

International Journal of Applied Engineering & Technology

# MACHINE LEARNING FOR PREDICTING STUDENT PERFORMANCE

**Khushilal Jatav*[1], Dr. Arpana Bharani[2] and Dr. Ritesh Yadav[3]**
[1,2]Department of Computer Science, Dr. A.P.J. Abdul Kalam University, Indore
[3]Department of Physics, Dr. A.P.J. abdul Kalam University, Indore
[1]dr.khushilal@gmail.com

## ABSTRACT
*There is a tremendous amount of computerized information generated in the area of data mining methods. One of the most important applications of Education Data Mining is the construction of models for predicting student accomplishment in academic institutions. Using their grades from their 10th, 12th, and preceding semesters, a formula has been devised to forecast their future success. Binomial logical regression, Decision tree, and Entropy and KNN classifiers are used to analyse the research. This framework would help the student realize their final grade and enhance their academic behaviour in order to get a better score. The aim of study machine learning for predicting student performance. As a result of these students' difficulties, these institutions should establish student assistance strategies. From 2009 through 2021, the relevant EDM literature on student dropouts and students at risk has been analysed in this systematic review. Students at danger of dropping out and students who have already dropped out were identified as potential targets for a variety of ML approaches, according to the findings of the study. Most research employ two datasets: student records from colleges and universities and online learning platforms. The importance of ML approaches in predicting at-risk pupils and dropout rates has been shown, and this has led to an increase in student achievement.*

*Keyword: Education data mining; Machine learning; MOOC; Student performance; Prediction*

## INTRODUCTION
Predicting pupils' performance has gotten increasingly challenging due to the abundance of data in educational databases. An established framework for measuring and monitoring students' achievement is not presently being explored. For the most part, there are two main explanations for this kind of occurrence. When it comes to forecasting student success, there is currently a lack of study on current ways of predicting student performance at institutions. The lack of investigation into the individual courses is the second problem. An overview of artificial intelligence systems used to predict academic achievement is the true purpose. Predictive algorithms are also being used to identify the most important qualities in student data. It is possible to increase student performance and advance more quickly and effectively using educational machine learning approaches. There may be advantages for students, teachers, and academic institutions, as well as an effect on them.

## LITERATURE REVIEW
**Ghassen Ben Brahim (2022)** There has been a great deal of research on predicting students' academic achievement throughout the course of their academic careers. Using this data, schools can make better judgments and implement changes that will lead to improved student outcomes. Electronic learning has been more popular in the post COVID-19 epidemic period, which has led to an increase in the availability of online learning data. ML-based models have been developed to predict students' success in online classrooms because of this. Digital electronics education and design suite data was used in this research to predict student success throughout a series of online interactive sessions. Text editing, the number of keystrokes, time spent in each activity, as well as the test score earned every session are tracked by this dataset. A total of 86 new statistical variables were extracted and semantically grouped into three broad groups based on distinct criteria in our proposed prediction model: (1) activity type, (2) timing statistics, and (3) the number of peripheral activity. To further limit this number of traits, we selected just the most important ones for training purposes. Using our suggested ML model, we want to forecast whether a student's performance would be poor or high. Random forest (RF), support vector machine, Nave Bayes, logistic regression, and multilayer perceptron (MLP) were all employed in our research. One 80/20

random data split was utilised for training and testing, two five-fold cross-validation tests were run on our model, and three training sessions were used for testing. Our model's classification accuracy using the RF classifier was the best, with a score of 97.4%. We showed that our model outperformed previous research that used a comparable experimental setting..

**M. Vahdat (2015)** Learning Analytics may be used to better analyse students' interactions with technology-enhanced learning (TEL) systems. In this study, we demonstrate that students' interaction data may be used to get insight into their learning processes. Our findings are based on data gathered from six lab sessions at the University of Genoa, where first-year Computer Engineering students used a digital electronics simulator. Students' learning processes are investigated and compared using Process Mining techniques. We use a complexity index to assess the usability of their process models. We then compare the academic accomplishments of the different groups of pupils. There is a positive association between students' grades at the end of the semester and their complexity, and a negative correlation between their difficulty in the lab sessions. Because of this, the degree of complication in process models may be used to predict the variety of learning routes taken by students.

**Nikola Tomasevic (2019)** To better understand and enhance the learning process, educational data mining has gained relevance and momentum as a result of a recent rise in data availability. Students who are at a "high risk" of dropping out of a course and predicting their future achievements, such as their final exam scores, are the focus of this paper. Its goal is to provide an in-depth analysis of and comparison between state-of-the-art machine learning techniques applied to the task of student exam performance prediction. With artificial neural networks, the most accurate forecasts were made by providing student involvement and historical performance data, whereas demographic information had no meaningful impact on the accuracy of the predictions. It was determined that suitable data gathering capabilities and student participation with the learning environment are prerequisites to ensuring significant amounts of data for analysis in order to fully leverage the student test performance prediction potential.

**Arto Hellas (2018)** Predicting how a class or program's students will do opens up new possibilities for improving educational results. It is possible for teachers to more properly manage time and resources when using effective performance prediction tools. Data mining is a field in which researchers strive to find traits and techniques that may be utilised for prediction, as well as measures of student achievement. In addition, research into student performance prediction aims to understand the fundamental causes of why certain characteristics function better than others and to find the linked features. Predicting student performance is the focus of this working group's literature review. According to our findings, both the volume and range of research being conducted in this field are both steadily rising. A number of study quality flaws were also found as a consequence of the review, necessitating more efforts in validation and replication by the scientific community.

**Mushtaq Hussain (2019)** To help instructors prevent students from dropping out before the final examinations and identify kids who need further aid, researchers are studying the prediction of student performance. Studying students' challenges in a future digital design course session is the goal of this research. Data collected by a TEL system known as the Digital Electronics Education and Design Suite (DEEDS) was subjected to machine learning algorithms for analysis. Artificial neural networks (ANNs), support vector machines (SVMs), logistic regression, Nave bayes classifiers, and decision trees were among the machine learning methods. Students may use the DEEDS system to go through a variety of digital design challenges while tracking their progress. Study variables included average duration, total number of activities, average idle time, average number of keystrokes and total associated activity for each exercise in the digital design course during individual sessions; study outcomes included grades for each session's students. On the basis of the data from the previous session, we developed and evaluated machine learning algorithms. Models were evaluated by performing k-fold cross-validation and calculating the receiver operating characteristic and root mean square error. It has been found that ANNs and SVMs are more accurate than other methods. This may be done with ease in the TEL system, thus educators should see an increase in student performance in the next session.

## METHODOLOGY

A research approach used to conduct a systematic literature review must be fair and thorough in order to assess all the existing research in the particular topic. For our systematic literature review, we used Okoli's guidelines. However, despite the fact that Kitchenham and many other researchers presented a full technique systematic literature review, most of them focused on just a small portion of the process, and only a few followed the whole process. An established approach for a systematic literature review is introduced via this method. But although though this study is primarily concerned with information systems, academics from any social science discipline may benefit greatly from it. Okoli's approach for systematic literature review is shown in Figure 1.

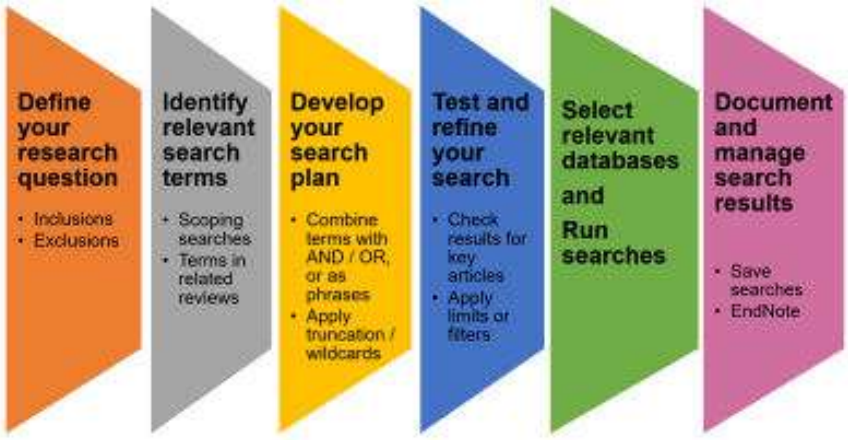


**Figure 1:** Okoli's guide for conducting a standalone Systematic Literature Review.

SLR reviewers are expected to identify and address the following research topics during the review process:

### 2.1. Research Questions

- What kinds of issues have researchers encountered when attempting to predict student performance in academic settings?
- What are the recommended remedies to these issues?
- Is there a lot of progress being made in this area of study?

### 2.2. Data Sources

For this study, we used six research databases to obtain the original data and to search for relevant publications that were related to our aims. Table 1 lists all of the databases that were used throughout the study process. Students at risk and their dropout rates between 2009 and 2021 were studied in depth utilising queries based on ML methods in these sources. Only the most relevant research articles were retained from the pre-determined searches, which resulted in numerous publications being manually screened out.

| Identifiers | Databases | Access Date |
|---|---|---|
| Sr.1 | ResearchGate | 4 February 2021 |
| Sr.2 | IEEE Xplore Digital Library | 4 February 2021 |
| Sr.3 | Springer Link | 6 February 2021 |
| Sr.4 | Association for Computing Machinery | 4 February 2021 |
| Sr.5 | Scopus | 4 February 2021 |
| Sr.6 | Directory of Open Access Journals | 4 February 2021 |

**Table1.DataSources.**

### 2.3. Used Search Terms

According to our study questions, the following search phrases (one by one) were used:

- Electronic Data Management (EDM) or Performance or eLearning or Machine Learning.
- Education Data Mining OR Prediction of Student Performance OR Evaluation of Students OR Analysis of Student Performance OR Prediction of Learning Curves.
- For example, "Students Intervention," "Dropout Prediction," "Student's dangers," "Students monitoring," "Requirements of students," and so on.
- *Predict* and machine learning as well as students.

### 2.4. The Paper Selection Procedure for Review

Finding, screening, eligibility verification, and completing inclusion requirements are all part of the paper selection process. The research articles were acquired separately by the authors, and they agreed on which publications to include. Using Okli's systematic review guide, the review selection approach is shown in detail in Figure 2.

### 2.5. Inclusion and Exclusion Criteria

#### 2.5.1. Inclusion

- Research on how to predict a student's performance;
- Accepted and published research papers in peer-reviewed journals or conferences that have undergone a blind review process;
- Documents produced between 2009 and 2021;
- Materials written in English.

#### 2.5.2. Exclusion

- Other studies employing ML that are not related to Student's Performance Prediction..
- Experiments or validation of planned procedures that have not been carried out.
- Editorials, Business Posters, Patents and longer versions of previously reviewed studies that have already been done, as well as technical reports and Wikipedia articles, are all examples of short papers..

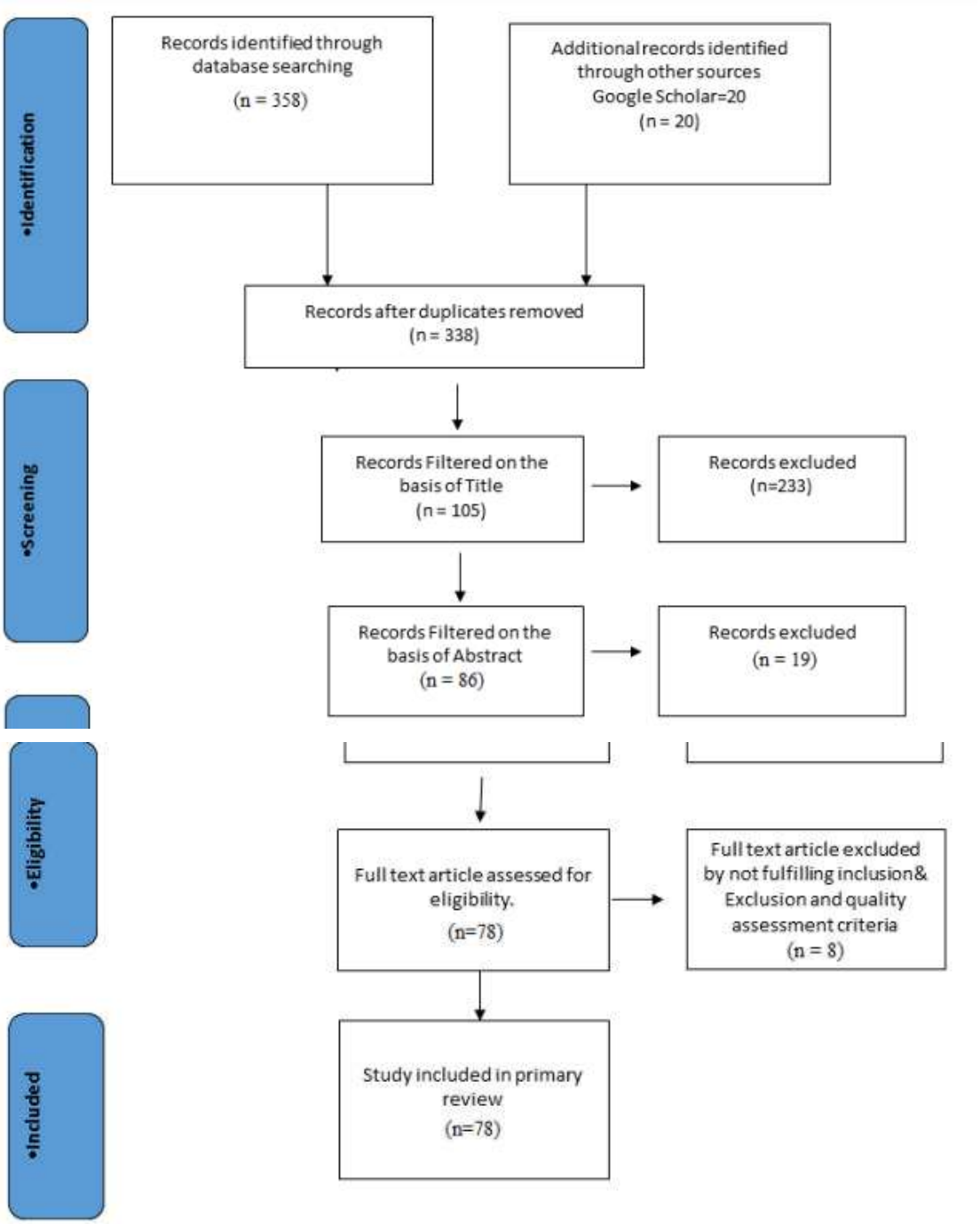


**Figure 2:** Detailed structure of review selection procedure after applying Okli's [12] for systematic review.

### 2.6. Selection Execution

The results of the search are used to compile a list of studies for further review. A bibliography tool is in charge of keeping track of all of the study's citations**.**

## RESULTS AND DISCUSSION

PISA 2005 test scores are being examined using machine learning and statistical methodologies. Several countries, including Germany, the United States, the UK, Spain and Italy as well as Japan and Canada were examined by the author in his study of PISA 2005 data. Studying how students and academic institutions interact to impact their academic performance was the goal of this project, which used a two-step method. Regression trees are used in the first stage to identify student-level factors associated with academic performance, which are nested inside schools. In the second stage, school value-additions are calculated using regression tree and boosting approaches, allowing for the identification of school-related attributes. The total number of characteristics at the school and student level were 19 and 18, respectively, in the PISA 2015 dataset from the nine nations. The sample size is the number of pupils (Table 2). The findings indicated that both student and school variables had an influence on kids' academic performance.

**Table 2:** Student sample size in the selected nine countries based on 2015 PISA dataset

| No. | Country | Sample Size |
|---|---|---|
| 1 | Canada | 20,058 |
| 2 | Australia | 14,530 |
| 3 | UK | 14,157 |
| 4 | Italy | 11,586 |
| 5 | Spain | 6736 |
| 6 | Japan | 6647 |
| 7 | Germany | 6504 |
| 8 | France | 6108 |
| 9 | US | 5712 |

An e-learning platform used four ways to accurately estimate student performance and identify students at risk: I prediction of academic achievement, identifying students at risk and determining the problems in an e-learning platform. Sixteen studies conducted between 2009 and 2021 suggest that student academic performance forecasting is an important research focus area for these techniques. A total of 12 studies were conducted throughout the same time period with the goal of identifying students who were at high risk for academic failure. Methodology and type characteristics utilised to establish applicable classification algorithms are distinct for each study. The most sought-after characteristic was the degree to which first-year students interacted with the e-learning platform. Only five research looked at how the e-learning platform affected students' performance and how it may be improved. DT, LR, NB, MT, and SVM were the most widely used algorithms. Table 3 contains information that may be used to forecast student performance and identify those who are at risk.

**Table 3:** Prediction of student performance and identification of students at risk in e-learning.

| Approach | Methodology | Attributes | Algorithms | Count |
|---|---|---|---|---|
| Performance prediction | Early prediction- ML | Socio-demographic | Rule- base | 2 |
| | Incremental ensemble | Teaching effectiveness | NB, 1-NN, and WINDOW | 2 |
| | Recommender system | Student's platform interaction | MT, NN, LR, LWLR, SVM, NB, DT, MLP | 5 |
| | Automatic measurement | Students' activity log | WATWIN | 2 |
| | Dynamic approach | 1st-year students | LR-SEQ, LR-SIM, DT, Rule-based & NB | 3 |
| | Semi-supervised ML | Secondary schools | YATSI, SVM, ANN | 6 |
| Identification of students at-risk | ML framework | At-risk of failing to graduate | SVM, RF, LR, Adaboost, CART, and DT | 3 |
| | Reducing feature set size | Early prediction of at-risk students | CART, C4.5, MLP, NB, KNN & SMO CF | 1 |
| | Student previous grades | Final GPA results | MF, RBM, GBDT, KNN, SVM, RF, | 3 |
| | Predictive models- grading | Identification of students at risk | DT, LDA, Adaboost | 2 |
| | Factors affecting- at- risk | Fast Learner, Average & Slow Learner | LR, SVM, DT, MLP, NB, and KNN Extra Tree (ET), RTV-SVM | 3 |
| Predict the difficulties of the learning platform | Examination of ML methods | Difficulties encountered on the e-learning system | ANN, LR, SVM, NBC, and DT | 2 |
| Performance of classifiers | Cross comparison | Comparison between five ML-based classifiers | NB, BN, ID3, J48, and NN | 2 |
| Evaluation of MOOC in developed countries | Discriminants of the PISA 2005 test score | Characteristics of students and academic institutions | ML and statistical methods | 1 |

Identifying kids who may be at risk of dropping out of school is crucial to determine what corrective steps should be taken. Methods used to determine dropout characteristics include curriculum and student performance, retention rate, dropout factor and early prediction. The traits and academic achievement of the student were often employed by studies to identify dropout features. Use of dynamic and static data sets helped to identify students who could drop out earlier. For dropout prediction, the most often used algorithms were DT, SVM, CART, KNN, and NB (Table 4).

**Table 4:** Prediction of student dropout using ML techniques and EDM methods.

| Approach | Attributes | Algorithms | Count |
|---|---|---|---|
| Features for dropout prediction including temporal features | students' personal characteristics and academic performance | DT, LR, SVM, ARTMAP, RF, CART, and NB | 10 |
| Curriculum-based and student performance-based features | Students performance class imbalance issues | K-NN, SMOTE | 2 |
| Retention rate | Freshman students | DT, Artificial Neural Networks (ANN) | 2 |
| Dropout factors | Evaluation of useful temporal models (Hidden Markov Model (HMM) | RNN combined with LSTM | 3 |
| Early-stage prediction of possible student dropout | pre-college entry information, and transcript information | ICRM2 with SVM, NB, DT, ID3, DL, and KNN, CART, and Adabooting Tree | 4 |

Predicting student performance and spotting at-risk kids early on is critical to preventing dropouts and devising effective interventions. Students' reading, quiz scores, and activity logs from the e-learning system were utilised in a total of 15 research investigations (Table 5). Static datasets were employed in nine investigations, whereas

dynamic and static datasets were used in a total of 14 studies. This shows that the learning platform's student performances and activities give a lot of input required to anticipate performance. Most often used methods for early prediction utilising static and dynamic data include KNN (kernel net), SVM (support vector machine), DT (data tree), radio frequency (RF) and infrared (ID3).

**Table 5:** Utility of static and dynamic data for early prediction of student performance.

| Approach | Attributes | Algorithms |
|---|---|---|
| Dynamic | Student performance data, student reading and quiz activity | K-NN, SMOTE, BM, SVM, NB, BN, DT, CR, ADTree, J48, and RF |
| Static | Enrolment and demographic data | Item Response Theory (IRT), ICRM2, SVM, NB, DT, ID3, DL, and KNN, CART and Adabooting Tree |
| Both | Pre-college entry information, and transcript information | ICRM2 with SVM, DL, ID3, KNN, DT, LR, SVM, ARTMAP, RF, CART, and NB |

## CONCLUSION

The research examines how machine learning may be used to track the academic progress of students. Decision trees, Entropy and KNN classifiers are utilised in the study of Binomial logical regression. This approach may assist the teacher in determining the student's academic performance and devising a more effective strategy for increasing their grades. To improve our dataset's accuracy, we want to add more characteristics in the future.. There were, however, just a handful of studies that suggested ways to provide students, teachers, and educators timely feedback to solve the issues. Research in the future will focus more on developing an efficient ensemble method to implement the ML-based performance prediction methodology in practise and searching for dynamic ways or methods to predict students' performance and provide automatic needed remedial actions to help students as early on as possible, Last but not least, we highlight the possible paths for future study employing ML approaches in forecasting students' performance. We'd want to put some of the great ideas already out there into practise while also emphasising the fluidity of student performance. With the help of these new insights, teachers will be better able to design effective interventions for their students and attain their educational goals with more accuracy.

## REFERENCES

[1] Brahim, G.B. Predicting Student Performance from Online Engagement Activities Using Novel Statistical Features. Arab J Sci Eng (2022). https://doi.org/10.1007/s13369-021-06548-w

[2] Vahdat, M.; Oneto, L.; Anguita, D.; Funk, M.; Rauterberg, M.: A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator. In: Design for Teaching and Learning in a Networked World, pp. 352–366. Springer, Cham (2015)

[3] Tomasevic, N.; Gvozdenovic, N.; Vranes, S.: An overview and comparison of supervised data mining techniques for student exam performance prediction. Comput. Educ. 143, 103676 (2020)

[4] Hellas, A.; Ihantola, P.; Petersen; A.; Ajanovski, V.; Gutica, M.; Hynninen, T; Liao, S.N.: Predicting academic performance: a systematic literature review. In: Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, pp. 175–199 (2018)

[5] Hussain, M.; Zhu, W.; Zhang, W.; Abidi, S.M.R.; Ali, S.: Using machine learning to predict student difficulties from learning session data. Artif. Intell. Rev. 52(1), 381–407 (2019)

[6] Buenaño-Fernández, D.; Gil, D.; Luján-Mora, S.: Application of machine learning in predicting performance for computer engineering students: a case study. Sustainability 11(10), 2833 (2019)

[7] Ofori, F.; Maina, E.; Gitonga, R.: Using machine learning algorithms to predict students performance and improve learning outcome: a literature based review. J. Inf. Technol. 4(1), 33–55 (2020)

[8] Huang, S.; Fang, N.: Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models. Comput. Educ. 61, 133–145 (2013)

[9] Rastrollo-Guerrero, J.L.; Gomez-Pulido, J.A.; Duran-Dominguez, A.: Analyzing and predicting students' performance by means of machine learning: a review. Appl. Sci. 10(3), 1042 (2020)

[10] Sundar, P.P.: A comparative study for predicting students academic performance using Bayesian network classifiers. IOSR J. Eng. IOSRJEN e-ISSN, 2250-3021 (2013)

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

[12] Ma, X.; Zhou, Z.: Student pass rates prediction using optimized support vector machine and decision tree. In: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), pp. 209–215. IEEE (2018)

[13] Masci, C.; Johnes, G.; Agasisti, T.: Student and school performance across countries: a machine learning approach. Eur. J. Oper. Res. 269(3), 1072–1085 (2018)

[14] Pardo, A.; Han, F.; Ellis, R.A.: Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. IEEE Trans. Learn. Technol. 10(1), 82–92 (2016)

[15] Gray, G.; McGuinness, C.; Owende, P.: An application of classification models to predict learner progression in tertiary education. In: 2014 IEEE International Advance Computing Conference (IACC), pp. 549–554. IEEE (2014)