Study Of Students' Performance Prediction Models Using Machine Learning

[1]Mr. S. Viswanathan, [2]Dr. S. Vengatesh Kumar

[1]Research Scholar, Department of Computer Science, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore. viswa.sankar77@gmail.com.
[2]Associate Professor, Department of Computer Applications(PG), Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.,gowthamvenkyy@gmail.com.

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract - Educational Data Mining (EDM) is a novel concept associated with developing methods for exploring the specific types of data produced by educational settings and using those approaches to effectively understand students and the environments in which they learn. Prediction attempts to shape trends that will allow it to predict results or learning outcomes based on available data. Predicting student success has become an appealing challenge for researchers. They develop an understandable and efficient model using supervised and unsupervised EDM techniques. This assists decision-makers in improving student performance. The task of deciding the best model leads to the emergence of various techniques from both EDM techniques. The numerous research models used to solve the problem of student success prediction using educational data mining are discussed in this paper. The primary purpose of this paper is to explain the methodology for implementing the proposed solution for student performance prediction, as well as to present the findings of a study aimed at evaluating the performance of various data mining classification algorithms on the given dataset in order to assess their potential usefulness for achieving the goal and objectives.

Keywords: - Educational Data Mining (EDM), Student Performance, Classification, Prediction, and Ensemble Model

1. Introduction

Data mining techniques are efficient research methods for finding hidden trends in massive, unintelligible datasets [2]. These methods are successfully used in a variety of fields, including accounting, pharmacy, human diseases, communications, healthcare, and education [3]. Educational data mining (EDM) is one of the research areas that uses data mining techniques on educational datasets to obtain useful information that can be interpreted and understood for decision making [4]. There are a number of variables that influence the decision to forecast and track student success in educational institutions. Educational institutions' decision-makers use these variables to create or build strategies to enhance and track students' academic performance [5].

The growing focus on data mining in the educational sector has resulted in the creation of a novel emerging research area known as Educational Data Mining (EDM). It is concerned with the study of educational data produced by educational environments [3]. Classification, Regression, Time Series Analysis, Clustering, Connection Rule Mining, and Neural Networks are some of the data mining methods widely used in EDM.

The aim of this paper is to evaluate educational data derived from students' transcripts and forecast their success so that they can take appropriate measures at the appropriate times to meet their expectations. Early prediction of student success in the correct manner would enhance both student retention and the assessment methods used by the students. This approach would also support educators and education officials by providing them with more knowledge about their students' learning abilities as well as how to better assist students who are falling behind in a given set.

Furthermore, evaluating students' learning and making assumptions about potential aspects of their success is important for an educational system in order to provide customized learning opportunities tailored to each student's particular needs or even to guide them to pursue technical education. Thus, it is critical to closely track students' performance in order to recognize potential retardation and to interfere proactively in their academic enhancement through the assignment of extra learning content, small group training, and so on.

1.1. Factors that Influence Academic Performance of Students

Education's primary aim has always been to boost student academic performance. Many studies have been performed over the years by researchers and educators to assess the factors that affect (positively or negatively) student achievement in their academic track. It was claimed that measuring student academic performance would

be difficult because student performance would be influenced by socioeconomic, psychological, and environmental factors.

Examinations hold a special role as a way of measuring a student's academic performance. In reality, students' exam performance is heavily influenced by three factors: demographics, academic climate, and socioeconomic factors.

- **Demographic Factors**

Demographic variables are personal characteristics such as age, gender, body mass index (BMI), food habits, and disabilities, as well as the form of family system, living area, and sibling structure.

- **Academic Environment factors**

Academic environmental variables are those that have a direct effect on a student's academic success at the higher education level and include continuous evaluation assessments, the type of education, the type of institute, the location of the institute, the medium of instruction, private tutoring, the marks earned at the final level, the type of community chosen at the higher education level, and extra-curricular activities.

- **Socio-Economic factors**

The socioeconomic status of a family is determined by family income, parental education level, parental occupation, and social status in the community. Families with a high socioeconomic status are also more active in preparing their young children for school since they usually have access to a wide variety of opportunities to encourage and facilitate the growth of young children. They are able to offer high-quality child care, books, and toys to their young children in order to enable them to engage in different learning experiences at home. They will have easy access to knowledge about their children's wellbeing, as well as their children's social, emotional, and cognitive development.

This study aims to explore previous studies on designing models to predict student performance in an educational environment. The review work was approached in a systematic manner by the authors of this paper. This strategy is intended to support the study's goals, which are:

- To define current prediction methods and tools used to predict student performance.
- To investigate and classify the type of variable used for the predictive process.
- To classify and investigate the researchers who used these learning models to assess student performance.

As a result, the ability to predict students' output with high precision at several stages of the huge quantity of dataset school period is regarded as important not only for educators and educational institutes, but also for students. More broadly, ‘‘knowledge exploration" will help educators better conduct their classes, recognize learning disabilities, and develop their teaching methods, while students can be given a preliminary assessment of their development and probably improve their performance.

The following is how the paper is structured: Section 2 addresses several similar works and explains why a new approach is needed. Section 3 describes our methodology steps and how the process will contribute to the creation of our novel prediction model. Section 4 investigates our experimental environments, findings, and discussions, accompanied by Section 5's conclusions and future work.

2. Related work

Predicting student academic performance is becoming an important factor in enhancing the standard of academic instruction, assisting students as they study, and providing tutors with more choices while preparing their students. Many works on this topic have been published in recent years. Here discovered several literature reviews that analyzed student academic performance modelling from various perspectives.

Ghorbani, R., & Ghousi, R. (2020) [5] Attempts to compare various resampling strategies, such as Borderline SMOTE, Random Over Sampler, SMOTE, SMOTE-ENN, SVM-SMOTE, and SMOTE-Tomek, to manage the imbalanced data issue when predicting students' success using two different datasets. Furthermore, the distinction between multiclass and binary classification, as well as feature structure, was investigated. Random Forest, K-Nearest-Neighbor, Artificial Neural Network, XG-boost, Support Vector Machine (Radial Basis Function), Decision Tree, Logistic Regression, and Nave Bayes are among the machine learning classifiers used in this analysis. As model validation techniques, the Random hold-out and Shuffle 5-fold cross-validation methods are used. The findings obtained using various evaluation criteria suggest that models would perform better with fewer classes and nominal features.

Almasri, A., et al. (2020) [6] proposed a coherent structure for developing a new supervised cluster-based (CB) classifier model The unified structure implements a clustering technique to arrange student historical records into a series of homogeneous clusters. Then, for each cluster, a classifier model is constructed, and the final unified classifiers, along with the centroids for each cluster, are used as the CB classifier model. According to the experimental findings, the CB model achieves a high accuracy performance of 96.25 percent. Furthermore,

employ feature selection techniques to pick appropriate features from a space of features. The model achieves a high accuracy output using relevant features, achieving 96.96 percent, with relevant features accounting for 57.4 percent of total features on average.

Popescu, E., & Leon, F. (2018) [7] proposed the method consists of using an advanced regression algorithm known as "Large Margin Nearest Neighbor Regression" (LMNNR) for grade prediction based on students' behavior on wiki, forum, and micro blogging tools. The findings are excellent, outperforming widely used regression algorithms. The key goal is to forecast students' success based on their social media footprints. Data is gathered from a Web Applications Design course in which students use wiki, forum, and microblogging resources for communication and collaboration in a project-based learning scenario. The research involves a total of 343 students from six consecutive course installments. For grade prediction, an advanced regression algorithm is used in addition to the novel settings and performance indicators. Excellent correlation coefficients are obtained, and 85% of predictions are within one point of the actual grade, outperforming traditional regression algorithms.

Chen, W., et al. (2018) [8] focus on Short courses with a single result assigned by the teacher at the end. Due to a lack of performance data and relatively limited enrollments, learner activity recorded as they engage with course content and with one another in Social Learning Networks (SLN) is critical for prediction. Based on the processing of behaviors collected on the modes of (human) learning in a course, our approach describes several (machine) learning features that are then used in acceptable classifiers. Through estimation on data captured from three two-week courses hosted through our delivery platforms, make three key observations: (i) behavioral data contains signals predictive of learning outcomes in short-courses (with classifiers achieving AUCs ≥ 0:8 after the two weeks), (ii) early detection is possible within the first week (AUCs ≥ 0:7 with the first week of data), and (iii) the content features have an "earliest" detection capability, while the SLN features become the more predictive set over time as the network matures.

Lau, E. T., et al. (2019) [9] proposed an approach that combines traditional statistical analysis with neural network modeling/prediction of student results Traditional statistical evaluations are used to determine the variables that are likely to influence the students' results. The neural network is represented by 11 input variables, two hidden neuron layers, and one output layer. As the backpropagation training rule, the Levenberg–Marquardt algorithm is used. The error performance, regression, error histogram, uncertainty matrix, and region under the receiver operating characteristics curve are used to test the neural network model's performance. Overall, the neural network model has achieved a good prediction accuracy of 84.8%, along with limitations.

Lee, C. S., et al. (2018) [10] proposed an For students' learning performance evaluation and educational applications, a particle swarm optimization (PSO) agent based on the Fuzzy Markup Language (FML) is proposed, and the proposed agent is based on data analysis from a traditional test and an item response theory (IRT)-based three-parameter logistic (3PL) model. Finally, the proposed work employs a K-fold cross validation process to assess the efficiency of the proposed agent. The experimental results show that the novel PFML learning mechanism performs well for parameter estimation and learning optimization. It is more complex in practice to overcome the application of adaptive evaluation agents. Furthermore, the agent must estimate students' skill item by item, which places a significant burden on the server, particularly for a group test. The proposed PFML will serve as a reference for educational research and pedagogy, as well as a critical co-learning framework for potential human–machine educational applications.

Xu, J., et al. (2017) [11] develops a novel machine learning approach for predicting student performance in degree programs that can address these main challenges has been developed. The suggested approach has two main characteristics. For making predictions based on students' changing performance states, a bilayered structure comprised of multiple base predictors and a cascade of ensemble predictors is first created. Second, a data-driven method focused on latent factor models and probabilistic matrix factorization is proposed to discover course relevance, which is critical for developing effective base predictors. The proposed study discusses the following challenges through detailed simulations on an undergraduate student dataset obtained over three years at UCLA: (1) Students vary considerably in terms of history and course selection; (2) Courses are not equally insightful for making accurate predictions; and (3) Students' changing progress must be factored into the forecast.

Several studies have been examined that explored, illustrated, and investigated the use of DM techniques in education. Several study studies that used EDM to analyze students' performance in higher education data have been analyzed. To summarize, several research are currently being performed to predict students' academic

performance in conventional classrooms or virtual education platforms. These studies yielded very interesting and logical results. Researchers have been involved in the study of prediction in academic competition.

3. PROBLEM DEFINITION

Predicting students' performance remains exceedingly difficult due to the large amount of data and the lack of an existing method to measure and study students' progress and performance. Their research also suggests that by using suitable data mining techniques (such as Decision Tree, Neural Networks, Nave Bayesian, K-Nearest Neighbor, and SVM), students' achievement and performance can be improved more effectively. The primary research issue is whether interactivity influences student output and how students implicitly interact with learning materials in various learning environments. Academic performance of students is a major factor in any educational institution; therefore, strategic programs in continuing inspiring or guiding the students for a better performance that may lead to a better future should be planned. Information derived from data in higher education institutions is not always accurate because the counting procedure does not always take other variables and attributes into account that could influence the extracted knowledge. Various educational data mining (EDM) techniques, such as classification, regression, and density estimation, have been used in the prediction of student results, such as forecasting student behaviors and correlating student interactivity.

Accurately predicting students' future performance dependent on their ongoing academic records is critical for carrying out required pedagogical measures to ensure students graduate on time and satisfactorily. As a result, innovative plans focused on information technology support must view student performance as part of the institutional culture. It is also important that the models used to assess student performance be applicable to other higher education institutions. Machine learning methods have proved to be useful for this function in recent years. The primary goal of this research is to create an integrated framework that enables machine learning to be used to predict student performance. Several machine learning methods, including decision trees, neural networks, support vector machines, and random forest, were used to assess the accuracy of student output predictions in the proposed work. The research results provide valuable insight into how to build a more effective and sensitive method to assist students in achieving their educational goals.

4. Proposed methodology

Predicting student performance is a critical challenge that is being investigated using EDM. This task anticipates the importance of an unspecified variable that defines the students in terms of outcome (Pass/Fail), grades, points, and so on. Predicting student attrition, errors, and progress are the key topics covered in this study's literature review. Every stakeholder in this domain desires an early warning system to forecast learning at an early stage. This early warning system lowered not only the cost of learning but also the amount of time and space required.

On regression or classification problems, different learning algorithms produce different results. Since the learning results from different algorithms vary, it is possible to boost the final prediction output for each algorithm, resulting in better results with the combined learning algorithms as compared to the results obtained with a single algorithm [22]. Ensemble learning is intended to improve predictive accuracy by combining predictions from multiple algorithms. Ensemble learning has been widely used in machine learning to boost efficiency by grouping individual algorithms on a variety of classification and regression tasks. Several ensemble strategies have been suggested, including voting, averaging, bagging, boosting, and piling. Stacking is a high-performance heterogeneous ensemble process. In recent years, it has been commonly used in data mining competitions. It can be thought of as a super multi-layer vision. Each layer contains one or more models, and the next layer of the model learns from the effects of the previous layer. Many models are used in machine learning to solve a binary classification problem, such as the regression algorithm, decision tree algorithm, kernel-based algorithm, Bayesian process algorithm, clustering algorithm, and so on. To boost robustness and generalization in a single model, stacking can be easily combined with different classifiers or regression models.

One of the most challenging problems is to improve the consistency of educational processes in order to improve student performance. Instructors should update their teaching methods to meet the needs of low-performing students and offer additional assistance to deserving students. The prediction results may assist students in developing a good understanding of how well or poorly they will perform in a course and then taking appropriate steps. Rising student retention is a long-term goal for all educational institutions worldwide. Increased retention has many positive results, including enhanced college credibility, ranking, and career opportunities for alumni, among others.

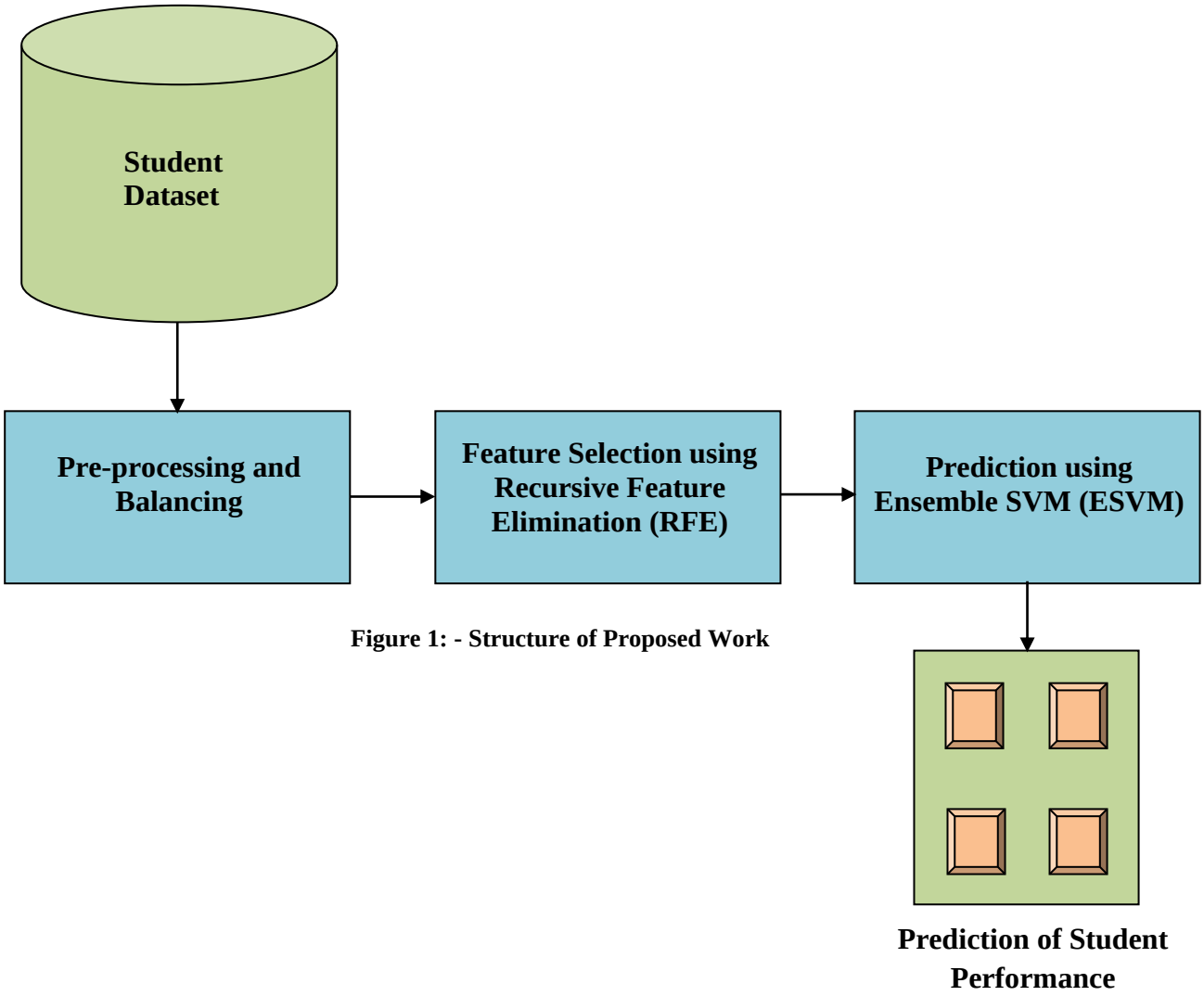

Figure 1: - Structure of Proposed Work

Predicting students' performance in postgraduate studies is crucial for any educational institution. It is particularly important for those who want to give students opportunities to do something useful in their field of study, as well as those who want to effectively manage the required teaching tools for excellent learning experiences. Furthermore, understanding students' performance in and course ahead of time is a must in order to assist at-risk students by minimizing the difficulties they face in their learning journeys and assisting them in excelling in the learning process.

Phase 1: - Pre-processing and Balancing

Data pre-processing is an important step in Machine Learning because the quality of data and the valuable knowledge that can be extracted from it directly affects our model's ability to learn; thus, it is critical that pre-process our data before feeding it into our model. Its aim is to transform raw data into a format that mining algorithms can use. During this process, the following tasks are completed.

- Data Integration.
- Data Cleaning.
- Discretization.

After data pre-processing, the data balancing approach is used to solve the class imbalance problem. The class imbalanced problem occurs when the number of instances in one class is significantly less than the number of instances in another class or classes. The proposed work assumes that the Cross-Validation approach is used to estimate the test error associated with a model in order to evaluate its accuracy.

Phase 2: - Feature Selection using RFE

Many attributes in the student output dataset may be inappropriate for classification purposes. The aim of feature selection is to choose an acceptable subset of features that can effectively represent the input data, thus reducing the dimensionality of the feature space and eliminating irrelevant data. The filter method searches for the fewest number of important features while ignoring the rest. It ranks the features using variable ranking techniques, with the highest ranked features being selected and added to the learning algorithm. The proposed work is an introduction. RFE, or Recursive Feature Elimination, is a common feature selection algorithm. RFE is common because it is simple to set up and use, and it is efficient at defining the features (columns) in a training dataset that are more or more important in predicting the target variable.

Phase 3: - Ensemble Learning Paradigm using Ensemble SVM (ESVM)

Ensemble approaches are strategies for creating several models and then integrating them to achieve better performance. Ensemble models usually yield more reliable results than a single model. This was the case in a number of machine learning competitions where the winning solutions employed ensemble methods. In the final step, implement the ensemble support vector machine (ESVM) classification technique, which was built by combining multiple diversity structures of SVM classifiers and thus has high generalization efficiency and classification precision. The proposed SVM ensemble learning model is made up of two different SVM classifier structures and five different kernel functions. The diversity of the ensemble members, in particular, is based primarily on different kernel function options and the structure of the SVM classifiers.

Every educational institute nowadays requires an effective student academic performance prediction model. However, resolving data quality problems in student success prediction models is often the most difficult task. This study proposed a model for predicting student performance dependent on the supervised learning technique ensemble SVM.

5. Performance evaluation

Evaluating classifier efficiency is an important part of comparing and selecting the best one. There are several methods for measuring and evaluating the output of machine learning algorithms. This paper employs a variety of assessment methods, including prediction Accuracy, Sensitivity, Precision, and F1-score; additionally, a statistical evaluation technique is employed for more reliable and efficient analyzing and comparing. Analyzing and evaluating the output of the classifiers is an essential technique. While assessment methods are easy to use, the results obtained can be misleading. Seeking the right model or system based on their strengths is therefore a crucial challenge. Evaluating classifier efficiency is an important part of comparing and selecting the best one. There are several methods for measuring and evaluating the output of machine learning algorithms. There are five commonly used different measures for evaluating classification consistency. Details are as follows:

- **CCI (Correctly Classified Instances)**: represents the number of correctly identified instances divided by the total number of instances Precision is a term that is commonly used.
- **ICI (Incorrectly Classified Instances)**: represents the number of instances that were wrongly labelled divided by the total number of instances.
- **Precision**: of algorithm represents the percentage of accurate classified instances from all truly classified instances.
- **Recall**: reflects the division number of correctly classified instances by the total number of all instances.
- **F-Measure**: measured from recall and precision values.

6. Conclusion

The primary aim of research is to significantly predict student performance in order to enhance academic outcomes. This can be done by the use of various educational data mining techniques to provide high-quality education. One way to achieve the highest degree of quality in the higher education sector is by accurate estimation of students' learning in educational institutions. There are numerous prediction models available using different mining techniques. Existing policies have largely been unable to respond to the increasing demands for higher and master training as mandated by the education framework. The current models are reviewed in this paper, and a novel model is proposed to effectively predict student success. This research work aims to specify the challenges and opportunities of quality education in higher education institutions, as well as provide a model for improving education quality.

References

1. Owadally, I.; Zhou, F.; Otunba, R.; Lin, J.;Wright, D.: Time series data mining with an application to the measurement of underwriting cycles. N. Am. Actuar. J. **23**, 1–16 (2019).
2. Hussain, S.; Dahan, N.A.; Ba-Alwib, F.M.; Ribata, N.: Educational data mining and analysis of students' academic performance usingWEKA. Indones. J. Electr. Eng. Comput. Sci. **9**(2), 447–459 (2018b).
3. Jauhari, F.; Supianto, A.: Building student's performance decision tree classifier using boosting algorithm. Indones. J. Electr. Eng. Comput. Sci. **14**(3), 1298–1304 (2019).
4. Sana, B.; Siddiqui, I.F.; Arain, Q.A.:Analyzing students' academic performance through educational data mining. 3c Tecnología:glosas de innovación aplicadas a la pyme **8**(29), 402–421 (2019).
5. Ghorbani, R., & Ghousi, R. (2020). Comparing different resampling methods in predicting Students' performance using machine learning techniques. *IEEE Access*, *8*, 67899-67911.
6. Almasri, A., Alkhawaldeh, R. S., & Çelebi, E. (2020). Clustering-Based EMT Model for Predicting Student Performance. *Arabian Journal for Science and Engineering*, 1-12.
7. Popescu, E., & Leon, F. (2018). Predicting academic performance based on learner traces in a social learning environment. *IEEE Access*, *6*, 72774-72785.
8. Chen, W., Brinton, C. G., Cao, D., Mason-Singh, A., Lu, C., & Chiang, M. (2018). Early detection prediction of learning outcomes in online short-courses via learning behaviors. *IEEE Transactions on Learning Technologies*, *12*(1), 44-58.
9. Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences*, *1*(9), 1-10.
10. Lee, C. S., Wang, M. H., Wang, C. S., Teytaud, O., Liu, J., Lin, S. W., & Hung, P. H. (2018). PSO-based fuzzy markup language for student learning performance evaluation and educational application. *IEEE Transactions on Fuzzy Systems*, *26*(5), 2618-2633.
11. Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, *11*(5), 742-753.
12. Altabrawee, H., Ali, O. A. J., & Ajmi, S. Q. (2019). Predicting Students' Performance Using Machine Learning Techniques. *JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences*, *27*(1), 194-205.
13. Bydžovská, H. (2016). A Comparative Analysis of Techniques for Predicting Student Performance. *International Educational Data Mining Society*.
14. 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.
15. Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education*.
16. Cazarez, R. L. U., & Martin, C. L. (2018). Neural Networks for predicting student performance in online education. *IEEE Latin America Transactions*, *16*(7), 2053-2060.
17. Tadayon, M., & Pottie, G. J. (2020). Predicting student performance in an educational game using a hidden markov model. *IEEE Transactions on Education*, *63*(4), 299-304.
18. Verma, K., Singh, A., & Verma, P. (2016). A review on predicting student performance using data mining method. *Futuristic Trends in Engineering, Science, Humanities, and Technology (FTESHT-16)*, 124-129.
19. Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, *13*(1), 61-72.
20. Kostopoulos, G., Kotsiantis, S., & Pintelas, P. (2015). Predicting student performance in distance higher education using semi-supervised techniques. In *Model and data engineering* (pp. 259-270). Springer, Cham.
21. Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015, July). Predicting students performance in educational data mining. In *2015 International Symposium on Educational Technology (ISET)* (pp. 125-128). IEEE.
22. Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting Students' performance In Distance Learning Using Machine Learning Techniques. *Applied Artificial Intelligence*, *18*(5), 411-426.
23. Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., & Ragos, O. (2020). Transfer learning from deep neural networks for predicting student performance. *Applied Sciences*, *10*(6), 2145.