Semiconductor Optoelectronics ISSN: 1001-5868

A NOVEL APPROACH FOR ANALYSIS AND PREDICTION OF STUDENTS ACADEMIC PERFORMANCE USING MACHINE LEARNING ALGORUTHMS

[1]Mr. S. Viswanathan, [2]Dr. S. Vengatesh Kumar.
[1]Research Scholar, Department of Computer Science, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.
allmyresearchpaper@gmail.com.
[2]Associate Professor, Department of Computer Applications (PG), Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.
gowthamvenkyy@gmail.com.

Abstract---Educational data mining has become an efective tool for exploring the hidden relationships in educational data and predicting students' academic performance. The prediction of student academic performance has drawn considerable attention in education. However, although the learning outcomes are believed to improve learning and teaching, prognosticating the attainment of student outcomes remains underexplored. To achieve qualitative education standard, several attempts have been made to predict the performance of the student, but the prediction accuracy is not acceptable. The main purpose of this research is significantly predict the student performance to improve the academic results. In order to accomplish the prediction with supplementary exactness, XGBoost based methods have been adopted. This work introduces a novel hybrid Lion-Wolf optimization algorithm to solve the problem of feature selection. Two level overlap improves the exploitation part. First phase overlap is used for feature selection and second phase used for adding some more important information and improve the classification accuracy. The XGBoost classifier improved the classification accuracy which is most famous classifier based on wrapper method. XGboost model using two different parameter adjustment methods are compared. XGBoost based on hybrid Lion-Wolf optimization performs better than traditional XGBoost on training accuracy and efficiency. Experiments are applied using the dataset and results prove that proposed algorithm outperform and provide better results.

Keywords— Educational Data Mining, Knowledge Discovery, Student Performance Prediction, Optimization Algorithm, Prediction, and Classification.

1. INTRODUCTION

EDM has emerged as a state of art research area in recent years. Student performance prediction model is one of the oldest and significant applications provided by EDM. Performance prediction model rely on different factors given by [1]. Many of factors are considered for EDM. Some of them are society factors, school factors, college factors, individual factor, family factors etc., however the factors considered of prediction are not consistent [2]. For effective prediction of student performance model, all factors related to student must be included based on the work results of [3].

Student performance prediction is a crucial job due to the large volume of data in educational databases. A lot of data has become available describing student's online behaviour and student engagement [4]. Online data has been used by a great number of researchers. Online learning and teaching is making a significant impact on the fabric of basic education. Predicting the performance of a student is a great concern to the basic education management. The scope of this paper is to identify the factors influencing the performance of students in different grades and to find out the best machine learning model to predicting student academic performance and helps to identify students with poor grades and then be evaluated and provided with new materials and methods to improve their results.

Early prediction is a new phenomenon that includes assessment methods to support students by proposing appropriate corrective strategies and policies in this field [5]. The main objective of this paper is to predict the students academic performsance based on the various factors. The final exam scores of college students reflect the students' learning effects to some extent, but the evaluation of learning effects cannot only be based on absolute scores. The traditional absolute score has certain limitations in reflecting the learning situation. The reasons are that difficulty of different courses is different, the marking standards of different teachers in the same course are different, and so on. In order to ensure the quality of talents, colleges and universities should not only judge the students by scores, but also analyze the learning effects of students, predict the academic performance of students in the future based on the analyzed results, and then set academicwarnings in time. This work will not only help colleges and universities to improve the quality of education, but also help students improve their overall performance, thereby improving the management of educational resources.

The research problem of this paper is to objectively evaluate students' academic performance from the perspective of features, and predict the future performance based on the existing performance. This study will be using various Machine Learning methods to predict student's academic performance. XGBoost with optimization technique are built and compared with traditional classification algorithms and making use of their predictive accuracy on the given data samples to predict student academic performance. Data is collected on KAGGLE and we will be focusing on student's engagement, how often they check their announcements, number of raised hands, number of accessed forum and number of accessed resources to predict student success. The process of predicting student performance using online logging's data is performed using various data mining techniques.

The further sections of the paper are structured as follows. Section 2 describes the previous research carried out in prediction and classification of student performance prediction. Section 3 describes the proposed XGBoost with optimization model in which the model components and its working procedure along with the algorithm are also presented in subsection 3.1 and 3.2. Section 4 discusses the performance evaluation of the proposed method and comparison with the existing model followed by section 6 that presents the conclusion of the proposed work.

2. LITERATURE SURVEY

Several methods and approaches were considered in predicting student performance;

most of these approaches are statistical in nature and designed for machine learning (ML) models. The models attempt to estimate an inherent correlation between input variables and identify patterns within the input data.

Student performance prediction models have targeted several metrics which are both quantitative and qualitative in nature. The amount of research work to predict quantitative metrics outweighs those for qualitative metrics [6]. Qualitative metrics havemainly focused on Pass/Fail or LetterGrade classifications of students in particular courses [7] or overall student assessment prediction in terms of high/average/low. By contrast, quantitative metrics have mainly attempted to predict scores or course/exam/assignment grades [8], range of course/exam/assignment grades [9], major dropout/retention rates [10], prediction of the time needed for exam completion, prediction of on-duration/delay of graduation and student engagement as well [11]. By contrast, quantitative metrics have mainly attempted to predict scores or course/exam/assignment grades [8], range of course/exam/assignment grades [9], major dropout/retention rates [10], prediction of the time needed for exam completion, prediction of on-duration/delay of graduation and student engagement as well [11].

Kim, B. H., et al., (2018) analyzed about the problem of student performance prediction - where a machine forecasts the future performance of students as they interact with online coursework. Reliable early-stage predictions of a student's future performance could be critical to facilitate timely educational interventions during a course. However, very few prior studies have explored this problem from a deep learning perspective. The authors proposed a new deep learning based algorithm, termed GritNet, which builds upon the bidirectional long short term memory (BLSTM). The results, from real Udacity students' graduation predictions, show that the GritNet not only consistently outperforms the standard logistic-regression based method, but that improvements are substantially pronounced in the first few weeks when accurate predictions are most challenging.

Al-Shehri, H., (2017) suggested for a mechanism for forecasting the performance of students can be useful in taking early precautions, instant actions, or selecting a student that is fit for a certain task. The need to explore better models to achieve better performance cannot be overemphasized. Due to the low result performance given by k-Nearest Neighbour algorithm, the author combined the k-NN and SVM on the dataset to predict the student's grade and then compared their accuracy. Empirical studies outcome indicated that Support Vector Machine achieved slightly better results. But combining KNN is called Lazy Learner (Instance based learning). It does not learn anything in the training period and SVM It does not execute very well when the data set has more sound i.e. target classes are overlapping.

Tripathi, A., et. al., (2019) provide a solution for the problem of student performance prediction. In the previous technique approach of SVM classifier is applied for the student performance prediction. The author proposed the approach called naïve bayes for the above mentioned problem. The results of proposed model are compared with existing model in terms of accuracy and execution time. But the Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

Priya, S., et. al., (2021) presented a framework for the problem of learnig outcomes of a student. The author proposed a new educational data mining framework in the form of rule-based recommender system. The proposed framework examines various factors in and around the institution. The framework is effective in identifying the student's weakness and making of relevant recommendations. Bur this framework is not suitable and tool for all the places because the factors are differ from one place to other.

Wang, Y., et. al., (2021) analyzed the problem of Student performance prediction. It is a critical research problem to understand the students' needs, present proper learning opportunities/resources, and develop the teaching quality. However, traditional machine learning methods fail to produce stable and accurate prediction results. The author proposed a a graph-based ensemble machine learning method that aims to improve the stability of single machine learning methods via the consensus of multiple methods. Ensembling is less interpretable, the output of the ensembled model is hard to predict.

Liu, D., et. al., (2020) aims to provide the solution for Student performance prediction. Usually every learning activity record has two types of feature data: student behavior and exercise features. The author proposed a novel framework for student performance prediction by making full use of both student behavior features and exercise features and combining the attention mechanism with the knowledge tracing model. Then, a fusion attention mechanism based on recurrent neural network architecture is used for student performance prediction.

From the analysis made from the literature, few of the methods are not suitable for handling the early prediction student performance. Some other methods find difficult to handle the students' performance prediction with various factors. Though few methods produce better result for both early prediction and factors used for predictions. Thus the proposed method exploits the usage of two components such as early prediction and features and factors chosen for students performance prediction commendably.

3. PROPOSED XGBOOST WITH OPTIMIZATION MODEL

The goal of this research is to predict the performance of students. This study contributes to the literature by predicting student academic performance and helps to identify students with poor grades can then be evaluated and provided with new materials and methods to improve their grades. Predicting students performance allow an instructor to spot non-engagement students based on their actions and activities from online learning platform. It also assists with identifying struggling students and giving teachers a proactive chance to come up with supplementary measures to improve their chances of passing during the course of their study programme.

From the literature that was reviewed various machine learning methods have been used by several researchers to predict student academic performance, they predicted student academic performance at the end of the study programme and they were unable able to detect which students may need immediate attention so that they lower the chances of them failing.

XGBoost stands for eXtreme Gradient Boosting [20]. It represents an instance of Gradient Boosting Machine (GBM) as a technique used mainly in the construction of both regression and classification predictive modeling problems. Experimentally, XGBoost is relatively faster than many other ensemble classifiers (such as AdaBoost). The impact of the XGBoost algorithm has been widely recognized in many machine learning and data mining challenges. In addition, it is a parallelizable algorithm, i.e. it can harness the power of multi-core computers, which also allows training on very large data sets.

Prediction of student academic performance has been carried out with different approaches. The purpose of the prediction model was to predict final performance of first year students based on their participation in online discussion forums. They have used multiple linear regression model and SVM classifier for prediction. They found that data collected for prediction was limited, so that generalization of prediction approach is not possible. Since, the prediction is done based on clustering approach. It is tedious and time consuming. Another regression method based prediction model was presented by [21]. The purpose of prediction model was to predict students marks in distant learning system. The prediction performance was better compared to conventional regression method but wide ranging conclusions seemed to be a problem.

3.1. Problem Definition

The problem formulation for student performance prediction model is reflected below. The utmost intention of proposed student performance prediction model is to predict semester marks of separate students based on data collected from various influencing factors related to students. Let X_i be data collected from S students based on family, schooling, environmental and individuality factors.The collected data is represented by $X_i = \{X_1;X_2; :::;X_N\}$. From collected data, best features for prediction are selected based on entropy measures. The entropy value for feature set is calculated and feature factors with minimal entropy measure are selected as best feature. Let $X_F = \{X_{F1} ;X_{F2} ; :::;X_{FM}\}$ be best feature selected for prediction. The feature sets are given as input to neural network to predict the semester performance. In XGBoost optimal weight for prediction is selected using proposed Lion-Wolf training algorithm the selected features are given to the XGBoost prediction.

3.2. Proposed Students Performance Prediction Model

In this section, detailed description about the proposed student performance prediction model is discussed. Primarily, data intended for prediction is collected from students.

we assume for a given data set: D = (x_i, y_i): i = 1, . . . , n, x_i ε R^m, y_i ε R, we have n observations with m features. Let ˆy_i be defined as the predict value by the model:

$$\widehat{Y_\iota} = \sum_{k=1}^{K} f_k\ (X_i)\,, f_k\ \in K \qquad ...equ.(1)$$

Chen and Guestrin [22] introduced the eXtreme Gradient Boosting (XGBoost) as described in equ.(2). At each iteration of gradient boosting, the residual will be manipulated to correct the previous predictor that the specified loss function can be optimized. The $f_k(x_i)$ corresponds to the prediction value given by the kth tree to the i^{th} sample. The set of functions f_k can be learned by minimizing the objective function:

$$Obj = \sum_{i=1}^{n} |\ \left(\widehat{Y_i},\ Y_i\right) + \sum_{k=1}^{K} \Omega\ (f_k) \qquad ...equ.(2)$$

With

$$\Omega\ (f_k) = \gamma^T\ \frac{1}{2}\ \lambda\omega^2 \qquad ...equ.(3)$$

The first term l in the equ.(2) is defined in the classification tree by $\left(\widehat{Y_\iota},\ Y_i\right) =$

$Y_i \mid n\,(1 + e^{-y_i}) + (1 - Y_i) \mid n\,(1 + e^{-y_i})$.. It represents the loss function and it measures the difference between the predicted value ˆyi and the target value y_i. While the second term V in the equ.(2) represents the regularization term; a factor used to measure the complexity of the tree f_k. Where g and l are the degrees of regularization. T and v. used in the equ.(3) are the numbers of leaves and the vector of values attributed to each leaf respectively.

3.2.1. Lion-Wolf Optimization Algorithm

In this section, proposed Lion-Wolf optimization algorithm for neural network training is deliberated. Lion-Wolf optimization algorithm is developed by integrating GreyWolf optimizer [23] and Lion optimization algorithm [24]. The proposed optimization algorithm is majorly based on grey wolf optimizer, but the position updating of grey wolf optimizer is hybridized with lion optimization algorithm increasing convergence rate and so avoiding the local optimal problem. Grey wolf optimizer is a meta heuristic optimization algorithm which is based on leader ship behaviors of grey wolf. The hunting behavior of grey wolf is used for optimization. The concept of prey hunting is extended in search space for best solution attainment which is nothing but the prey. In grey wolf optimizer, the best three fitness functions is called alpha, beta and delta wolfs. Remaining wolfs in population is called omega wolfs. The alpha, beta and delta wolf encircle the prey (solution) based on encircling behaviour and the remaining wolfs in population update position based on position of search agents. In this paper, the position updation of grey wolf optimizer is hybridized with lion optimization algorithm. Lion optimization algorithm is based on social behaviour of lions. The female lion updating of Lion algorithm is integrated into position updating equation of grey wolf optimizer. Generally, the local optima problem in Lion and GWO algorithm is low, by hybridization it is further reduced increasing convergence rate.

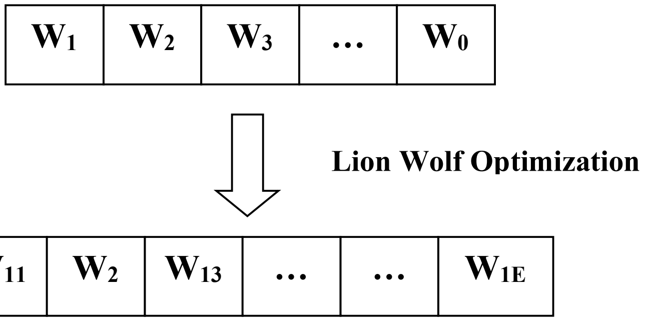

Fig.1: Hierarchy of Performance Influencing Factors

(a) Solution Encoding: - In proposed Lion-Wolf optimization algorithm, solution is grey wolf population. The possible solution is represented in solution encoding. The grey wolf population represents vectors of neuron weight for training. The length of solution i.e. number of grey wolfs in population is equal to number of weights required in training of XGBoost. Example for solution encoding is given in Fig. 1

(b) Fitness Evaluation: - The fitness function used in proposed Lion-Wolf Optimization training algorithm for XGBoost is deliberated in this section. The fitness evaluation function used is Mean Square Error. For each of solution in grey wolf population which is nothing but optimal vector of connection weight, training is performed in XGBoost. After training, the

Mean square error function is calculated between actual response and desired response of network. Mean Square Error function value is given in Eqn. 4

$$\mathrm{MSE}\left(\mathrm{W_{pq}}\right) = \frac{1}{\mathrm{mn}} \sum_{\mathrm{p}=1}^{\mathrm{n}} \sum_{\mathrm{q}}^{\mathrm{m}} \left(\mathrm{C_{pq}} - \mathrm{B_{pq}}\right) \qquad \text{... equ. (4)}$$

Where, n is number of input, m is number of output, C is desired response value of XGBoost and B is actual response of XGBoost for considered solution vector. For all solution in grey wolf population, MSE is calculated and solution with best fitness i.e. which has response equal to that of desired response is chosen optimal weight.

(c) Lion-Wolf Training Algorithm: - The steps involved in proposed Lion-Wolf optimization algorithm is discussed below; Each solution in grey wolf population represents solution vector containing weights for training of algorithm. Optimal weight for training which produce target output of XGBoost is intend. Optimization using Lion-Wolf Optimization algorithm selects the optimal weight. The position and solution updation of Lion-Wolf optimization algorithm updates solution vector over consecutive iterations.

Step 1: Parameter Initialization.
Step 2: Population Initialization
Step 3: Fitness Calculation.
Step 4:Lion Fusion.
Step 5: Lion-Wolf Hybridization.
Step 6: Position Updation.
Step 7: Solution Updation.
Step 8: Iteration.

4. RESULTS AND DISCUSSIONS

The experimentation of proposed student performance prediction model predicting student semester marks in college is performed in a personal computer with following specification; i) Windows 8 Operating System ii) Intel Core i-3 processor iii) 2GB Physical memory. The software tool used for implementation of proposed student performance prediction model is Weka in Java.

4.1. Performance Measures

4.1.1. Accuracy

Accuracy: represents one of the most commonly used metrics for rating performance. It estimates the proportion of correctly classified of students' answers by measuring a ratio between the answers correctly classified and the total number of answers..

Accuracy = TruePositives / (TruePositives + FalsePositives)

Table 1: - Comparision of Accuracy

Dataset	Existing Algorithm (Logistic Regression) (%)	Existing Algorithm (Standard XGBoost) (%)	Proposed Algorithm (XGBoost with Optimization) (%)
UCI - Student Performance Data Set	91.5	93	98
Kaggle - Student Performace Data Set	90.8	91.5	96.9

In Table 1, the results are reported for different feature selection methods for the different dataset. On classifying the dataset employing original features it is noted that the classification accuracy of the proposed algorithm XGBoost woth Optimization technique is greater than the existing logistic regression and standard XGBoost algorithm.

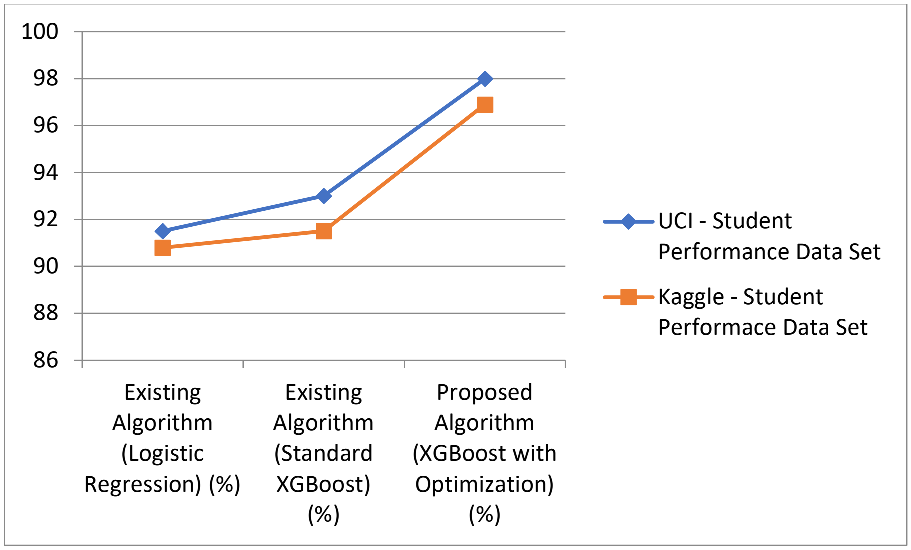

Fig.2: - Comparision of Accuracy with different Dataset

On applying the proposed XGBoost woth Optimization technique, the accuracy is increased significantly to 98 & 96.9%. The highest accuracy is reported for different dataset when the proposed XGBoost woth Optimization technique algorithm is employed. Figure 2, shows the better performance of classification rate for different dataset employing proposed XGBoost woth Optimization technique and existing Logistic Regression & standard XGBoost classifier with respect to accuracy.

5. CONCLUSION

Many researchers still have to use machine learning rather than deep learning algorithms while conducting knowledge-tracing experiments. For this purpose, the aim of conducting this research is to focus on the importance of exploiting technical improvements along with the pedagogical contributions that could be proposed. Thus, researchers could obtain higher predictive accuracy, if the use of one of the machine learning algorithms is indispensable. In this work, we have applied some of the most powerful Ensemble Learning methods, namely XGBoost with Optimization algorithm, to solve real-world classification problems in the educational system. The contribution aims to evaluate whether a single classifier can lead to sufficient optimization, and to specify to what extent the use of Ensemble Learning classifiers could improve the single KT classifier. The evaluation of our proposed models is done through a comparison against the original XGBoost and Logistic Regression model, as a more powerful Knowledge Tracing approach. For all datasets used, the results obtained have shown that Ensemble learning XGBoost with Optimization algorithm versions are more valuable compared to the XGBoost and Logistic Regression model. Moreover, the results have demonstrated that XGBoost with Optimization algorithm has the ability to predict student future acquisition with the highest performance. While, regardless of the remarkable improvement that can be provided by all ensemble learning models, it would be mentioning that the use of these classifiers requires certain environmental conditions in order to provide better results.

REFERENCES

[1] Touron, J. (1983). The determination of factors related to academic achievement in the university: Implications for the selection and counselling of students. Higher Education, 12(4), 399-410.

[2] Malvandi, S., & Farahi, A. (2015). Provide a method for increasing the efficiency of learning management systems using educational data mining. Indian Journal of Science and Technology, 8(28), 1.

[3] Shetgaonkar, P. R. (2015). Predicting the impact of different Variables on Students Academic Performance using Artificial Intelligence. Int. J. Comput. Sci. Inf. Technol, 6(2), 1367-1370.

[4] M. D. Dixson, “Measuring student engagement in the online course: the Online Student Engagement scale (OSE).(Section II: Faculty Attitudes and Student Engagement)(Report),” Online Learning Journal (OLJ), 19(4), 143, 2015, doi:10.3102/00346543074001059.

[5] Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human behavior, 104, 106189.

[6] Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students’ performance by means of machine learning: A review. Applied sciences, 10(3), 1042.

[7] Ma, X., & Zhou, Z. (2018, January). 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.

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

[9] Ofori, F., Maina, E., & Gitonga, R. (2020). Using machine learning algorithms to

predict studentsâ€™ performance and improve learning outcome: a literature based review. Journal of Information and Technology, 4(1).

[10] Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Computers & Electrical Engineering, 66, 541-556.

[11] Pardo, A., Han, F., & Ellis, R. A. (2016). Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. IEEE Transactions on Learning Technologies, 10(1), 82-92.

[12] Kim, B. H., Vizitei, E., & Ganapathi, V. (2018). GritNet: Student performance prediction with deep learning. arXiv preprint arXiv:1804.07405.

[13] Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., ... & Olatunji, S. O. (2017, April). Student performance prediction using support vector machine and k-nearest neighbor. In 2017 IEEE 30th canadian conference on electrical and computer engineering (CCECE) (pp. 1-4). IEEE.

[14] 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).

[15] Tripathi, A., Yadav, S., & Rajan, R. (2019, July). Naive Bayes classification model for the student performance prediction. In 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) (Vol. 1, pp. 1548-1553). IEEE.

[16] Priya, S., Ankit, T., & Divyansh, D. (2021). Student performance prediction using machine learning. In Advances in parallel computing technologies and applications (pp. 167-174). IOS Press.

[17] Asselman, A., Khaldi, M., & Aammou, S. (2021). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. Interactive Learning Environments, 1-20.

[18] Wang, Y., Ding, A., Guan, K., Wu, S., & Du, Y. (2021). Graph-based Ensemble Machine Learning for Student Performance Prediction. arXiv preprint arXiv:2112.07893.

[19] Liu, D., Zhang, Y., Zhang, J., Li, Q., Zhang, C., & Yin, Y. (2020). Multiple features fusion attention mechanism enhanced deep knowledge tracing for student performance prediction. IEEE Access, 8, 194894-194903.

[20] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[21] Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. Artificial Intelligence Review, 37(4), 331-344.

[22] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[23] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. Advances in

engineering software, 69, 46-61.

[24] Yazdani, M., & Jolai, F. (2016). Lion optimization algorithm (LOA): a nature-inspired metaheuristic algorithm. Journal of computational design and engineering, 3(1), 24-36.