Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Student Peformance Prediction Using Feature Imbalance Aware Xgboost Algorithm

Shashi Rekha H[1] , Dr. Chetana Prakash[2]

[1]Assistant Prof. Drpt Of Cse Vtu P-G Center.

[2]Prof. Cse Biet Devangiri.

Abstract—in the current situation of the pandemic, most of the schools and institutions have changed their mode of teaching to the online mode, hence there is a sudden increase in the e-Learning Systems. Due to the e-Learning Systems, many students are facing issues connecting to these platforms. Some of the issues include no electricity, no proper internet connection, etc. Hence, there is a slight decrease in the student's performance. Furthermore, some of the institutions are trying to improve the student performance and quality of education in the e-Learning Systems using Data Mining (DM) employed Machine Learning (ML) Technique. These techniques are used to analyze the student activity such as session time, login time, time spent in the e-Learning Systems, etc., and then predict the performance of the student. Some of the studies have shown that the Machine Learning-Based techniques give a correct result only when the data is balanced. Hence it is required to choose the correct Machine Learning algorithm according to the data. Most of the existing Student Performance Prediction Techniques have designed their models by combining various Machine Leaning Algorithms to choose the best model according to the data. Furthermore, these techniques have not incorporated the feature importance to predict the performance of the student. Hence, this results in poor performance mostly for the multi-label classification. Thus, this paper gives a model using the XGBoost (XGB) Algorithm, named Feature Imbalance Aware XGB (FIA-XGB). The FIA-XGB uses the effective cross-validation technique to learn the correlation between the features and increase the performance of the model efficiently. The results show better performance in terms of prediction accuracy when compared with the existing Machine Learning Student Performance Prediction models.

Keywords—Handover execution, Heterogeneous wireless network, Radio access technology selection, Machine learning.

I. INTRODUCTION

In the current situation of the pandemic, most of the educational institutions, corporate businesses have changed their way of working. Furthermore, as the Internet and Information Technology (IT) are gradually increasing in the current situation, most of the teaching is being done in the online mode called e-Learning platform [1] for most of the educational institutions. There are many challenges faced by the teachers and the e-Learning platforms to provide proper teaching methods, assess the students, etc. Hence there is a requirement for a model which predicts the performance of the student. Furthermore, there are many challenges to providing an accurate and reliable student performance prediction model [2]. Moreover, by designing an accurate model using the session streams acquired by the e-Learning platforms, we can predict the behavior of the student and student performance in the e-Leaning platforms. This is will help to increase the student's performance by giving the correct content to the student so that he can focus more on that content to improve his/her performance.

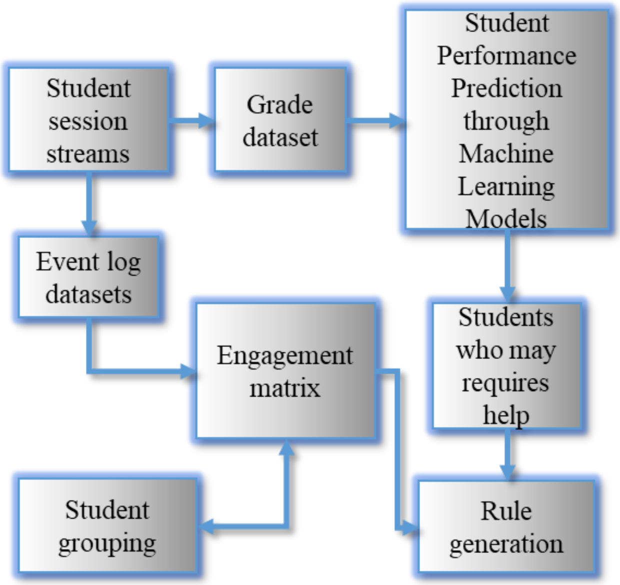

Fig. 1. General framework eLearning platform.

Attaining the correct content to improve student performance is a big challenge in the current situation [3]. Some of the methods such as the adaptive personalizing method have been used to understand the student profile [4], [5]. In recent times, Machine Learning (ML) and Data Mining (DM) techniques are being used for the prediction of the performance of the student. The DM methods can establish useful data from the e-Learning data stream sessions [6] as shown in Figure 1. Moreover, these methods have an improved decision-making performance [7] using the data [8], [9]. Both the ML [10], [11], [12] and DM [13] methods provide promising results and are widely used in the network security business and education [14], [15], [16]. A new technique to extract the data from the education has been emerging named Education Data Mining (EDM) [17] as shown in Fig. 1, to enhance the students learning style [18], to understand the student behavior [19], and also to improve the student performance [20]. The EDM method is combined with various kinds of information [21], such as student session stream data, student academic performance data, and administration data, etc. In [22], [23], they have provided an EDM dataset that has been collected from various E-learning systems and contains various databases. In this, they have used various machine learning models and an ensemble learning technique to predict the performance of the student. The results show that the model has better prediction accuracy when compared with the existing systems. Though when the dataset is an imbalance in nature, then this model has failed to give better results and hence has a poor classification accuracy. In addressing this paper presents feature imbalance aware-XGB for student performance prediction by incorporating effective cross validation mechanism.

Research Contribution are as follows:

- Improved better feature selection highly imbalanced student session stream data.
- The FIA-XGB achieves better prediction accuracy than existing student performance classification models.

Manuscript organization. In section II the detail survey of various existing methodologies and its limitation have been highlighted. The proposed methodology is discussed in section III. The experiment study using student session stream data is given in section IV. The last section significance of work is given and future research direction for enhancing student performance prediction outcomes.

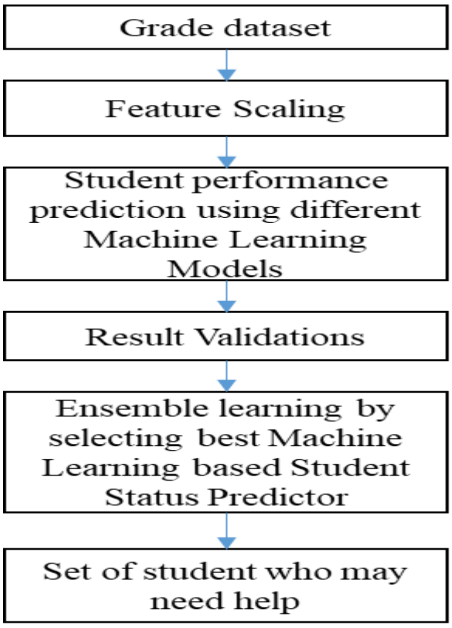
Grade dataset
Feature Scaling
Student performance prediction using different Machine Learning Models
Result Validations
Ensemble learning by selecting best Machine Learning based Student Status Predictor
Set of student who may need help

Fig. 2. Existing ensemble-based classifier for student performance prediction.

II. LITTERATURE SURVEY

This section conduct survey of recent work to enhance student performance in platform leveraging data mining and machine learning models; and highlight the limitation of recent works.

Author name	Methodology and significance	Limitations
Hussain et al., [5], 2018	The model focused on analyzing student session stream data of open university	However, when employed the model for different student session data these ML-

	through machine learning models. Good classification accuracy is achieved for Open University dataset.	based model performance badly.
Krishnamurthy et al., [8]	Designed Student performance and risk prediction, risk through feedback according to context-based cognitive skill ranks.	The model work only prior information of course is available and when tested under new environment poor classification accuracy is achieved [7].
Moubayed et al., [24]	Designed Student engagement level prediction in e-learning platform employing K-mean clustering algorithm.	The model does not provide good result when feature size are varied considering multi-class classification.
Injadat et al., [22], 2020	Designed ensemble-learning by combining multiple ML algorithm such as SVM, RF, NB, MLP, and KNN for predicting the student performance at early stages and halfway as shown in Fig. 2.	However, exhibit poor result when training dataset is imbalanced in nature.
Injadat et al.,	Modelled an optimized bagging ensemble learning	The model fails to establish feature impacting

[23], 2020	algorithm for improving prediction accuracy of student performance.	performance of classifier. Along with poor classification is incurred when data is imbalanced in nature.

III. Student Performance Prediction using Feature Imbalance Aware XGBoost Algorithm

This section present the working structure of proposed student prediction model using Data Imbalance XG Boost algorithm. XG Boost algorithm is an improvised version of Gradient Boosting algorithm [25] where weaker classifiers are combined together for constructing strong classifier for attaining better classification outcomes.

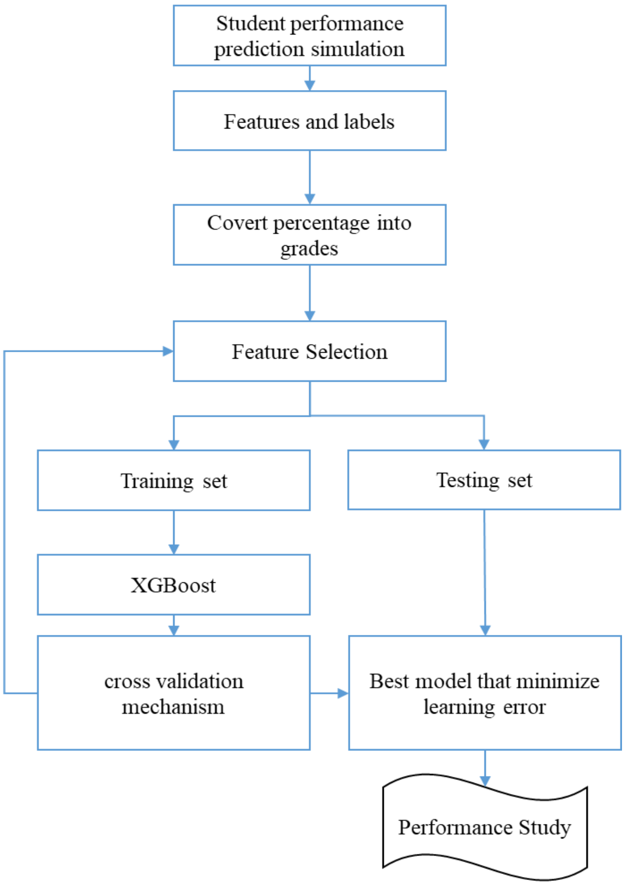
Student performance prediction simulation
Features and labels
Covert percentage into grades
Feature Selection
Training set
Testing set
XGBoost
cross validation mechanism
Best model that minimize learning error
Performance Study

Fig. 3. Architecture of proposed Feature Imbalance Aware XG Boost (FIA-XGB) algorithm for Student performance prediction.

Let consider a student session stream data $E = \{(y_j, z_j); j = 1 \ldots o, y_j \in \mathcal{S}^n, z_j \in \mathcal{S}\}$, which composed of o samples of data with n features. Let $\hat{z}_j$ defines the predicted outcome by models as follows

$$\hat{z}_j = \sum_{l-1}^{L} g_l(y_j), \quad g_l \in G \tag{1}$$

where g_l defines distinct regression tree and $g_l(y_j)$ defines respective prediction outcome provided by respective $l-\text{th}$ tree with respect to $j-\text{th}$ sample. The regression tree g_l and its function can be learned through minimization of following objective equation

$$\mathcal{O} = \sum_{j=1}^{o} m(z_j, \hat{z}_j) + \sum_{l=1}^{L} \beta(g_l) \tag{2}$$

In this work m defines training loss operation for measuring variance among predicated value $\hat{z}_j$ and the actual value z_j. In order to avoid over-fitting problem, the parameter β is used for penalizing complexity of predictive model as follows

$$\beta(g_l) = \delta U + \frac{1}{2}\mu||x||^2 \tag{3}$$

where δ and μ defines the regularization parameter, U defines the leaf size and x defines score of different leaf. The ensemble tree is constructed is through summation process. Let $\hat{z}_j^{(u)}$ defines the prediction outcome of the $j-\text{th}$ sample considering $u-\text{th}$ iterations, it requires to add $g_u$for minimizing the below defines functions

$$\mathcal{O}^{(u)} = \sum_{j=1}^{o} m\left(z_j, \hat{z}_j^{(u-1)} + g_u(y_j)\right) + \beta(g_l) \tag{4}$$

In this work the feature selection process of standard XGB is modified by establishing better feature importance outcome to achieves improved prediction scheme. The feature selection process is improved by optimizing the cross validation with minimal validation error as follows

$$\mathrm{CV}(\sigma) = \frac{1}{SM}\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{j\in G_{-k}} P\left(b_j, \hat{g}_\sigma^{-k(j)}(y_j, \sigma)\right) \tag{5}$$

In Eq. (5), to select ideal $\widehat{\sigma}$ for optimizing the student prediction model is attained as follows

$$\widehat{\sigma} = \underset{\sigma\in\{\sigma_1,\ldots,\sigma_l\}}{\arg\min}\ \mathrm{CV}_s(\sigma) \tag{6}$$

In Eq. (5), M defines size of training dataset considered, P(·) defines loss function, and $\hat{g}_\sigma^{-k(j)}(\cdot)$ defines a function to compute coefficients. The proposed FIA-XGB based student performance prediction model achieve better classification accuracy in comparison with ensemble-based classifier as shown in below section.

IV. Result and Analysis

This section discusses the result achieved using proposed FIA-XGB over existing ensemble-based student performance prediction model [22] in terms of accuracy, sensitivity, specificity, precision, and F-measure are used for validating student performance prediction model. The student session stream dataset used are obtained from [22]. The Fig. 4 shows the accuracy achieved using ensemble is 0.65 and FIA-XGB is 0.922. The Fig. 5 shows the specificity achieved using ensemble is 0.65 and FIA-XGB is 0.925. The Fig. 6 shows the sensitivity achieved using ensemble is 0.857 and FIA-XGB is 0.9448. The Fig. 7 shows the precision achieved using ensemble is 0.857 and FIA-XGB is 0.9509. The Fig. 8 shows the F-measure achieved using ensemble is 0.857 and FIA-XGB is 0.9473. From overall result achieved we can state the FIA-XGB is very efficient in comparison with ensemble based student performance prediction model.

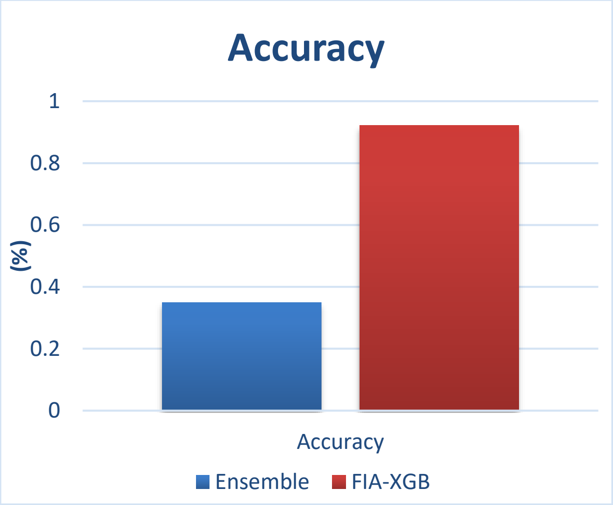

Fig. 4. Accuracy performance.

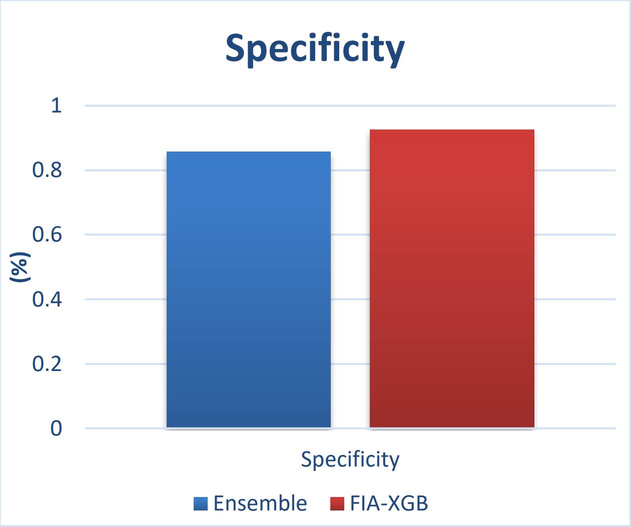

Fig. 5. Specificity performance.

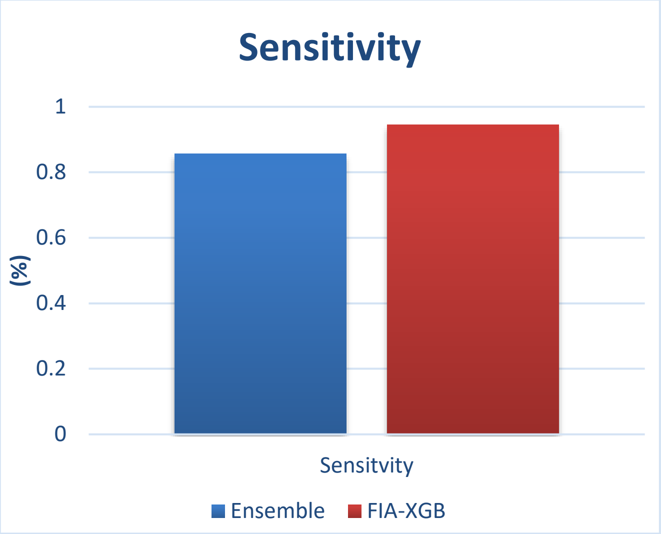

Fig. 6. Sensitivity performance.

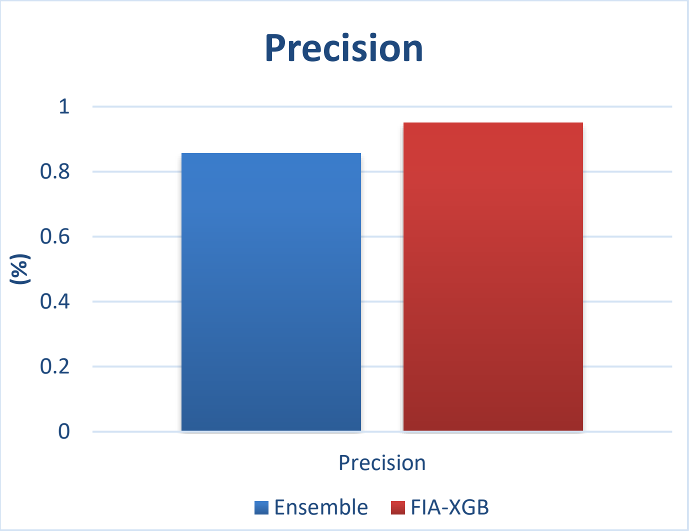

Fig. 7. Precision performance.

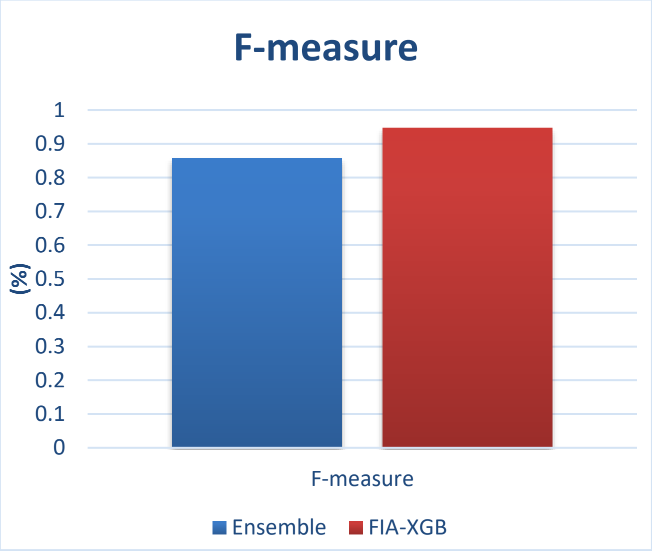

Fig. 8. F-measure performance.

V. Conclusion

This paper presented a feature imbalance aware XGB by incorporating effective cross-validation scheme that work well even when training data is imbalanced. Prediction enhancement is achieved through effective feature ranking mechanism. Experiment is conducted using standard student session stream data. The proposed FIA-XGB model significantly improves accuracy, sensitivity, specificity, precision, and F-measure performance in comparison with ensemble based student performance prediction model.

Future work would consider improving performance of FIA-XGB model under more diverse dataset.

References

[1] G. R. A. Moubayed, M. Injadat, A.B. Nassif, H. Lutfiyya, A. Shami, E-learning: Challenges and research opportunities using machine learning data analytics, IEEE Access 6 (2018) 39117–39138, http://dx.doi.org/10.1109/ACCESS.2018.

[2] F. Essalmi, L.J.B. Ayed, M. Jemni, S. Graf, Kinshuk, Generalized metrics for the analysis of e-learning personalization strategies, Comput. Hum. Behav. 48 (2015) 310–322, http://dx.doi.org/10.1016/j.chb.2014.12.050.

[3] J. Yang, J. Ma, S.K. Howard, Usage profiling from mobile applications: A case study of online activity for Australian primary schools, Knowl.-Based Syst. (2019) http://dx.doi.org/10.1016/j.knosys.2019.105214.

[4] Wakjira, A., & Bhattacharya, S. (2021). Predicting Student Engagement in the Online Learning Environment. International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 16(6), 1-21. http://doi.org/10.4018/IJWLTT.287095.

[5] Hussain, Mushtaq & Zhu, Wenhao & Zhang, Wu & Abidi, Raza. (2018). Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. Computational Intelligence and Neuroscience. 2018. 1-21. 10.1155/2018/6347186.

[6] G. Kaur, W. Singh, Prediction of student performance using weka tool, Int. J. Eng. Sci. 17 (2016) 8–16.

[7] Dhankhar A., Solanki K., Dalal S., Omdev (2021) Predicting Students Performance Using Educational Data Mining and Learning Analytics: A Systematic Literature Review. In: Raj J.S., Iliyasu A.M., Bestak R., Baig Z.A. (eds) Innovative Data Communication Technologies and Application. Lecture Notes on Data Engineering and Communications Technologies, vol 59. Springer, Singapore. https://doi.org/10.1007/978-981-15-9651-3_11

[8] MD, S., Krishnamoorthy, S. Student performance prediction, risk analysis, and feedback based on context-bound cognitive skill scores. Educ Inf Technol (2021). https://doi.org/10.1007/s10639-021-10738-2.

[9] Alyahyan, Eyman & Dustegor, Dilek. (2020). Predicting Academic Success in Higher Education Literature Review and Best Practices. International Journal of Educational Technology in Higher Education. 17. 10.1186/s41239-020-0177-7.

[10] M. Injadat, F. Salo, A.B. Nassif, A. Essex, A. Shami, Bayesian optimization with machine learning algorithms towards anomaly detection, in: 2018 IEEE Global Communications Conference, GLOBECOM, 2018, pp. 1–6, http://dx.doi.org/10.1109/GLOCOM.2018.8647714.

[11] L. Yang, A. Moubayed, I. Hamieh, A. Shami, Tree-based intelligent intrusion detection system in internet of vehicles, in: 2019 IEEE Global Communications Conference, GLOBECOM, 2019.

[12] A. Moubayed, M. Injadat, A. Shami, H. Lutfiyya, DNS typo-squatting domain detection: A data analytics & machine learning based approach, in: 2018 IEEE Global Communications Conference, GLOBECOM, IEEE, 2018, pp. 1–7.

[13] Namoun A, Alshanqiti A. Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. Applied Sciences. 2021; 11(1):237. https://doi.org/10.3390/app11010237

[14] Ayouni S, Hajjej F, Maddeh M, Al-Otaibi S (2021) A new ML-based approach to enhance student engagement in online environment. PLoS ONE 16(11): e0258788. https://doi.org/10.1371/journal.pone.0258788.

[15] S. M. Aslam, A. K. Jilani, J. Sultana and L. Almutairi, "Feature Evaluation of Emerging E-Learning Systems Using Machine Learning: An Extensive Survey," in IEEE Access, vol. 9, pp. 69573-69587, 2021, doi: 10.1109/ACCESS.2021.3077663.

[16] Khanal, S.S., Prasad, P., Alsadoon, A. et al. A systematic review: machine learning based recommendation systems for e-learning. Educ Inf Technol 25, 2635–2664 (2020). https://doi.org/10.1007/s10639-019-10063-9.

[17] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, D.J. Murray, Q. Long, Predicting academic performance by considering student heterogeneity, Knowl.-Based Syst. 161 (2018) 134–146, http://dx.doi.org/10.1016/j.knosys. 2018.07.042.

[18] Juhanak, L., Zounek, J., and Rohlíková, L. (2019). Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. Comput. Hum. Behav. 92, 496–506. doi: 10.1016/j.chb.2017.12.015.

[19] Liu, Q., Tong, S., Liu, C., Zhao, H., Chen, E., Ma, H., et al. (2019). "Exploiting cognitive structure for adaptive learning," in in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (Anchorage, AK), 627–635.

[20] Wang, F., Liu, Q., Chen, E., Huang, Z., Chen, Y., Yin, Y., et al. (2020). “Neural cognitive diagnosis for intelligent education systems,” in Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 6153–6161.

[21] J. Hung, B.E. Shelton, J. Yang, X. Du, Improving predictive modeling for at-risk student identification: A multistage approach, IEEE Trans. Learn. Technol. 12 (2) (2019) 148–157, http://dx.doi.org/10.1109/TLT.2019.2911072.

[22] Injadat, Mohammadnoor & Moubayed, Abdallah & Nassif, Ali & Shami, Abdallah. (2020). Systematic Ensemble Model Selection Approach for Educational Data Mining.

[23] Injadat, Mohammadnoor & Moubayed, Abdallah & Nassif, Ali & Shami, Abdallah. (2020). Multi-split Optimized Bagging Ensemble Model Selection for Multi-class Educational Data Mining.

[24] A. Moubayed, M. Injadat, A. Shami, H. Lutfiyya, Student engagement level in e-learning environment: Clustering using k-means, Am. J. Dist. Educ. (2020) http://dx.doi.org/10.1080/08923647.2020.1696140.

[25] Chen, Tianqi & Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System. 785-794. 10.1145/2939672.293978