

# Machine learning based education data mining through student session streams

Shashirekha Hanumanthappa<sup>1</sup>, Chetana Prakash<sup>2</sup>

<sup>1</sup>Department of Computer Science and Engineering, Visvesvaraya Technological University, Mysore, India

<sup>2</sup>Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere, India

---

## Article Info

### Article history:

Received Oct 16, 2023

Revised Dec 18, 2023

Accepted Dec 30, 2023

---

### Keywords:

Data imbalance

E-learning

Ensemble algorithm

Feature importance

Machine learning

---

## ABSTRACT

Recently, significant growth in using online-based learning stream (i.e., e-learning systems) have been seen due to pandemic such as COVID-19. Forecasting student performance has become a major task as an institution is focusing on improving the quality of education and students' performance. Data mining (DM) employing machine learning (ML) techniques have been employed in the e-learning platform for analyzing student session streams and predicting academic performance with good effects. A recent, study shows ML-based methodologies exhibit when data is imbalanced. In addressing ensemble learning by combining multiple ML algorithms for choosing the best model according to data. However, the existing ensemble-based model does not incorporate feature importance into the student performance prediction model. Thus, exhibits poor performance, especially for multi-label classification. In addressing this, this paper presents an improved ensemble learning mechanism by modifying the XGBoost algorithm, namely modified XGBoost (MXGB). The MXGB incorporates an effective cross-validation scheme that learns correlation among features more efficiently. The experiment outcome shows the proposed MXGB-based student performance prediction model achieves much better prediction accuracy contrary to the state-of-art ensemble-based student performance prediction model.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



---

## Corresponding Author:

Shashirekha Hanumanthappa

Department of Computer Science and Engineering, Visvesvaraya Technological University

Ring Road, Hanchya Sathagally Layout, Mysore, Karnataka 570019, India

Email: shashirekha\_h2k22@rediffmail.com

---

## 1. INTRODUCTION

With the wide usage of the internet and the growth of information technology have affected the way academics and industries learn i.e., it is moved from the conventional offline mode to online mode namely the e-learning platform [1]. Especially during the COVID-19 pandemic period, all classes have moved to an online model, highlighting the significance of the e-learning platform. However, significant challenges exist in providing a reliable and accurate model to predict student performance [2]. Designing an effective assessment model for understanding student behavior using session streams of the e-learning platform will aid in improving students' academic performance by providing personalized content.

Personalized content delivery for improving student performance according to individual behavior in the e-learning platform is the major challenge of the current century [3]. Adaptive personalizing techniques for understanding learner profiles have been emphasized [4], [5]. Recently, data mining (DM) and machine learning (ML) have been used for building student performance prediction models. The DM has been used for establishing useful insight from student session stream data of the e-learning platform as shown in Figure 1;

alongside, improves decisionmaking performance by establishing behavior patter from data [6]–[9]. Both ML and DM methodologies are very promising in different fields such as business, and network security including education. Recently, a new field has emerged namely education data mining (EDM) for enhancing learning style, understanding behavior, and improving student performance [10]–[13]. The EDM data is composed of different information such as administration data, student session stream activity, and student academic performance data. Here they provided an EDM dataset collected from different databases and e-learning systems. Here different ML models and an ensemble learning mechanism are constructed for predicting student performance during the course. The outcome shows ensemble model outperforms another model in terms of prediction accuracy [14]–[16]. However, when data is imbalanced these model fails to establish feature affecting the predictive model; thus, providing poor classification accuracies. The objective of this paper is to build an effective student prediction model for predicting student grades during the course through an ensemble-based ML model that works well for student session stream e-learning data [17]–[19]. Existing models construct ensemble learning by combining multiple ML models. However, these models are effective to address binary classification problems and when put forth under multi-label classification problems considering data imbalance, these methods exhibit poor accuracy [20], [21]. The aforementioned limitations motivate this research work to develop an improved student performance prediction model through improved ensemble methodology [22], [23]. This paper presents an effective student performance prediction through an improved ensemble-based ML model. First, the model briefs a detail of the ensemble algorithm namely XGBoost. Then, discusses the limitation of standard XGBoost when data is imbalanced. In addressing a modified XGBoost based student, a performance prediction model is presented [24], [25]. The modified XGBoost (MXGB) encompasses an improved cross-validation mechanism for establishing features affecting the accuracy of the student performance prediction model. Finally, an ensemble-based ML is constructed for building an effective student performance predictive model. Here research significance is discussed: i) the proposed student performance prediction model employs an efficient ensemble-based predictive model through MXGB, which works well even when data is imbalanced; and ii) the MXGB encompasses an improved cross-validation mechanism to study which feature impacts the accuracy of the student prediction model; and the proposed student performance prediction model achieves better receiver operating characteristic (ROC) performance such as accuracy, sensitivity, specificity, and sensitivity, precision, and F-measure comparison with the state-of-art student performance prediction model.

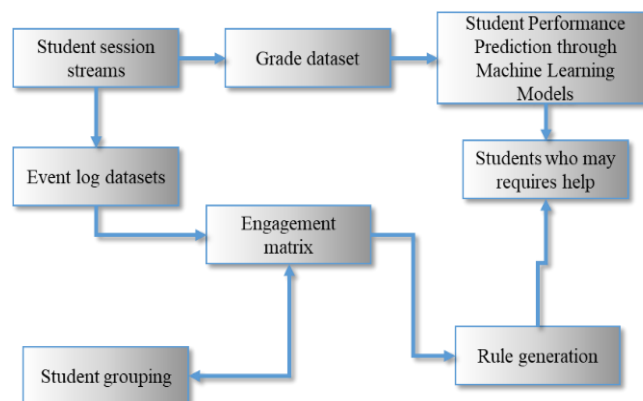


Figure 1. General design of student performance prediction through ML models

In section 2, ML model for EDM of student session streams. In section 3, the outcome was achieved using the proposed MXGB-based student performance prediction model over the existing ensemble-based existing proposed student performance prediction model. In the last section, the significance of the MXGB-based student performance prediction model over the existing ensemble-based student performance prediction model is discussed.

## 2. MACHINE LEARNING MODEL FOR EDM OF STUDENT SESSION STREAMS

This section presents an improved ML model namely MXGB for EDM of student session streams. The MXGB is an improvement of the standard XGBoost by considering an effective feature selection mechanism. The dataset of standard EDM is defined as (1):

$$E = \{(a_1, b_1), (a_2, b_2), \dots, (a_m, b_m)\} \quad (1)$$

where  $j=1,2,3, \dots, m$ , outlines row size considered,  $b_j \in \{-1,1\}$  defines  $j^{th}$  row output, and  $a_j$  defines  $n$ -dimension vector of self-determining features experimental of row  $j$ . In general, EDM data has diverse features that are multi-dimensional. Nonetheless, with fewer rows  $m$ . Thus, for studying and designing student performance prediction model  $\hat{G}$ , for forecasting the real estimation of actual  $G$  is defined as (2):

$$g: A \rightarrow B \quad (2)$$

in this work modifying the feature selection process during training XGBoost through minimization of the objective function and effective student performance prediction model is designed as shown in Figure 2.

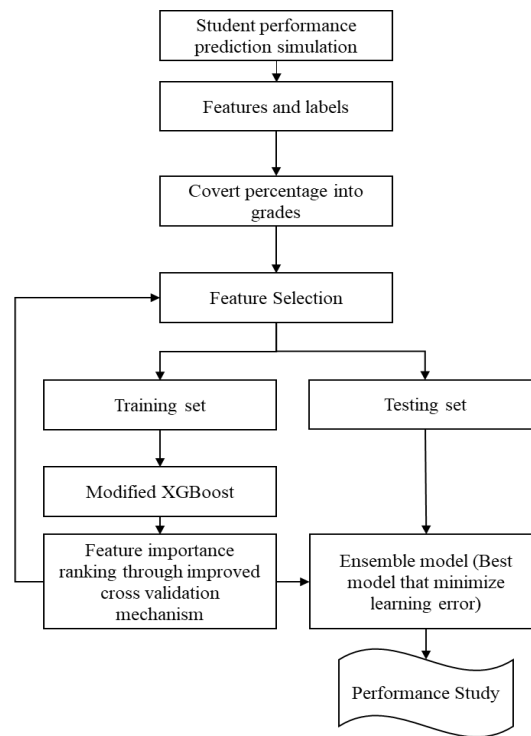


Figure 2. Proposed ML model for EDM of student session streams

### 2.1. XGBoost prediction algorithm

XGBoost algorithm is an improvised version of the gradient boosting algorithm [25] where weaker classifiers are combined for constructing strong classifiers for attaining better classification outcomes. Let consider a student session stream data  $E = \{(y_j, z_j); j = 1 \dots o, y_j \in S^n, z_j \in S\}$ , which composed of  $o$  samples of data with  $n$  features. Let  $z_j$  the predicted outcome by models as (3):

$$Z_j = \sum_{l=1}^L g_l(y_i), g_l \in G \quad (3)$$

where  $g_l$  defines a distinct regression tree and  $(y_i)$  defines the respective prediction outcome provided by the respective  $l$ -th tree concerning  $j$ -th sample. The regression tree  $g_l$  and its function can be learned through the minimization of the following objective in (4).

$$G = \sum_{j=1}^o m(z_j, z_j) + \sum_{l=1}^L \beta(g_l) \quad (4)$$

In this work,  $m$  defines training loss operation for measuring variance among predicated value  $z_j$  and the actual value  $z_j$ . To avoid the over-fitting problem, the parameter  $\beta$  is used for penalizing the complexity of the predictive model as (5):

$$\beta(g_l) = \delta U + \frac{1}{2}\mu\|x\|^2 \quad (5)$$

where  $\delta$  and  $\mu$  define the regularization parameter,  $U$  defines the leaf size and  $x$  defines the score of the different leaves. The ensemble tree is constructed is through a summation process. Let  $\hat{z}^{(u)}$  define the prediction outcome of the  $j$ -th sample considering  $u$ -th iterations, it requires to add  $g_u$  for minimizing the (6):

$$G^{(u)} = \sum_{j=1}^o m(z_j, \hat{z}_j^{(u-1)} + g(y)) + \beta(g) \quad (6)$$

the (6) is simplified by eliminating constant parameter through second-order Taylor expansion as (7):

$$G^{(u)} = \sum_{j=1}^o [h_j g_j(y_j) + \frac{1}{2} i_j g_u(y_j)^2] + \beta(g_l) \quad (7)$$

where  $h_j$  defines the first-order gradient concerning  $m$  as (8):

$$h_j = \partial \hat{z}_z^{(u-1)} m(z_j, \hat{z}^{(u-1)}) \quad (8)$$

where  $i_j$  defines the first-order gradient concerning  $m$  as (9):

$$i_j = \partial^2 \hat{z}_j^{(u-1)} m(z_j, \hat{z}_j^{(u-1)}) \quad (9)$$

therefore, the predictive model objective parameter is expressed using the (10).

$$G^{(u)} = \sum_{j=1}^o [h_j g_j(y_j) + \frac{1}{2} i_j g_u(y_j)^2] + \delta U + \frac{1}{2}\mu \sum_{k=1}^U x_k^2 \quad (10)$$

The simplified representation of the (10) is given as (11):

$$\mathcal{O}^{(u)} = \sum_{j=1}^U [(\sum_{j \in j_k} h_j) x_j \frac{1}{2} (\sum_{j \in j_k} i_j + \mu) x_k^2] + \delta U \quad (11)$$

where  $j_k$  defines the sample set of leaf  $k$ , which is represented as (12) and (13):

$$G^{(u)} = \sum_{j=1}^U [(\sum_{j \in j_k} h_j) x_j \frac{1}{2} (\sum_{j \in j_k} i_j + \mu) x_k^2] + \delta U \quad (12)$$

$$j_k = \{j | r(y_j = k)\} \quad (13)$$

where  $r$  defines the size of the tree, which is fixed, the optimal weights  $x_k^*$  of leaf  $j$  is obtained through the (14).

$$x_k^* = \frac{H_k}{I_k + \mu} \quad (14)$$

In addition, the respective optimal size is obtained as (15):

$$G^* = \frac{1}{2} \sum_{k=1}^U \frac{H_k^2}{I_k + \mu} + \delta U \quad (15)$$

where  $H_k$  is represented as (16):

$$H_k = \sum_{j \in j_k} h_j \quad (16)$$

similarly,  $I_k$  is represented as (17).

$$I_k = \sum_{j \in j_k} i_j \quad (17)$$

The  $G^*$  defines the qualities of tree  $r$  where a smaller value indicates better tree structure. Though XGBoost is efficient in obtaining high prediction accuracy; however, poor feature selection under unknown environments or when data is imbalanced exhibit degradation of prediction accuracy. In addressing the research problem, an effective feature selection within training data is modeled in the next sub-section.

## 2.2. Modified XGBoost prediction algorithm

In this work, the feature selection process of standard XGBoost is modified by establishing better feature importance outcomes to achieve an improved prediction scheme. The feature selection process is improved by optimizing the cross-validation with a minimal validation error. The K-fold cross-validation scheme is used for optimizing the outcome of the predictive model where the dataset is randomly divided into  $K$  subset of equal size. Then, for constructing the student prediction model  $K-1$  is used, and the remaining is used for optimizing the prediction error of the student prediction model. Lastly, the mean of the prediction error of different combinations.

$K$  is used for optimizing the cross-validation error. After that, a grid of  $l$  appropriate outcomes is obtained for obtaining optimal prediction that minimized cross-validation error considering feature importance, and the student prediction model with minimal cross-validation error is chosen. The proposed cross-validation scheme with effective feature selection is composed of two phases. In the first phase, the main feature is selected from feature subsets. In the second phase, features chosen from the first phase are utilized for constructing an effective student performance prediction model. The traditional single-fold cross-validation error is constructed as (18):

$$CV(\sigma) = \frac{1}{M} \sum_{k=1}^k \sum_{j \in G_k} P(b_j, \hat{g}_\sigma^{-k(j)}(y_j, \sigma)) \quad (18)$$

however, the above equation does not identify which feature affects the accuracy of the predictive model. In addressing this work an effective cross-validation with effective feature selection with high importance affecting prediction accuracy is modeled as (19):

$$CV(\sigma) = \frac{1}{SM} \sum_{s=1}^s \sum_{k=1}^k \sum_{j \in G_k} P(b_j, \hat{g}_\sigma^{-k(j)}(y_j, \sigma)) \quad (19)$$

in (19), selecting ideal  $\hat{\sigma}$  for optimizing the student prediction model is attained as (20).

$$\hat{\sigma} = \underset{\sigma \in \{\sigma_1, \dots, \sigma_2\}}{\operatorname{argmin}} CV_s(\sigma) \quad (20)$$

In (19),  $M$  defines the size of the training dataset considered,  $(\cdot)$  defines the loss function and  $\hat{g}_\sigma^{(j)}(\cdot)$  defines a function to compute coefficients. The (19) is executed iteratively for constructing the best student performance prediction model (i.e., its optimization of training error is done in the first phase; the parameter is passed onto the second phase to understand and update the feature importance characteristic into the predictive model. The optimization process to obtain effective features is obtained through the minimization process of objective function employing gradient decent mechanism. The effective feature is selected employing the ranking method  $(\cdot)$  for constructing a student performance prediction model through the (21):

$$r(a) = \begin{cases} 0 & \text{if } n_j \text{ is not selected} \\ 1 & \text{if } n_j \text{ is selected as optimal prediction model } j = 1, 2, 3, \dots, n \end{cases} \quad (21)$$

the feature subset is constructed as (22):

$$F_s = \{r(n_1), r(n_1), \dots, r(n_n)\}, \quad (22)$$

the ideal feature with maximum score considering varied  $K$ -folds instance is obtained as (23).

$$F_{sk} = \{r(n_1), r(n_1), \dots, r(n_n)\}, \quad (23)$$

Then, compute the number of occurrences a particular feature is selected for  $K$  feature subsets having maximum score and the final feature subset is obtained as (24):

$$F_{sfinal} = \{f_s(p_1), f_s(n_1), \dots, f_s(n_n)\}, \quad (24)$$

where  $(\cdot)$  depicts a case when where  $n^{th}$  feature is selected/not and mathematically represented as (25).

$$F_s(a) = \begin{cases} 0 & \text{if } q_j \text{ is chosen lesser than } \frac{K}{2} \text{ time, } j = 1, 2, 3, \dots, n \\ 1 & \text{if } q_j \text{ is chosen greater or equal to } \frac{K}{2} \text{ times, } j = 1, 2, 3, \dots, n \end{cases} \quad (25)$$

The aforementioned equation is used for the generation of a subset of  $n'$  selected features, where  $n^{th}$  describe how many times a feature is selected. The enterprise performance management (EPM) training data utilized is a subset through selected features for building an effective student prediction model. To reduce randomness during the training process,  $K$  -folds are built by iterating  $S$  number of times in the first phase. In the second phase, for reducing variance subset of features is selected. Therefore, the proposed MXGB-based student performance prediction model significantly improves overall prediction accuracy in comparison with state-of-art ML-based student performance prediction schemes.

### 3. RESULT AND ANALYSIS

In this section, student performance prediction using the proposed MXGB and other existing ML-based student prediction methods are studied [22]. The e-learning dataset from [22] is used for performance analysis. The selection of the dataset is based on a comparison paper [22]. The model is a ML model for performing student performance prediction implemented using the Python 3 frameworks. The ROC performance metrics such as accuracy, sensitivity, specificity, precision, and F-measure are used for validating the student performance prediction model. The accuracy is computed as (26):

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (26)$$

where  $TP$  defines true positive,  $FP$  defines false positive,  $TN$  defines true negative, and  $FN$  defines false negative. The sensitivity is computed as (27):

$$Sensitivity = \frac{TP}{TP+FN} \quad (27)$$

the specificity is computed as (28):

$$Specificity = \frac{TN}{TN+FP} \quad (28)$$

the precision is computed as (29):

$$Precision = \frac{TP}{TP+FP} \quad (29)$$

the F-measure is computed as (30).

$$F - \text{measure} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \quad (30)$$

#### 3.1. Predictive model performance evaluation

In this section different ML-based student, performance prediction model in terms of specificity and sensitivity is studied. Figure 3 shows the specificity outcome achieved using different student performance prediction models such as random forest (RF), logistic regression (LR), and ensemble-based [22]. XGBoost-based, and proposed MXGB-based. The RF-based attain a specificity of 0.875, the LR-based attain a specificity of 0.75, ensemble-based attain a specificity of 0.857. XGBoost-based attain a specificity of 0.8502, and the proposed MXGB-based attain a specificity of 0.946. A higher value of specificity i.e., closer to 1 is considered a good prediction model. Thus, the proposed MXGB-based student performance prediction model is much more efficient than other ML-based student performance prediction models in terms of specificity. Figure 4 shows the sensitivity outcome achieved using different student performance prediction models such as RF-based, LR-based, and ensemble-based. XGBoost-based, and proposed MXGB-based. The RF-based attains a sensitivity of 1, the LR-based attains a sensitivity of 0.857, ensemble-based attains a sensitivity of 0.857. XGBoost-based attain a sensitivity of 0.9449, and the proposed MXGB-based attain a sensitivity of 1. A higher value of sensitivity i.e., closer to 1 is considered a good prediction model. Thus, the

RF-based proposed MXGB-based student performance prediction model is much more efficient than other ML-based student performance prediction models in terms of sensitivity. However, the MXBG-based brings tradeoffs between higher sensitivity and specificity; thus, attaining much better student performance prediction accuracies.

Further, performance is validated considering different ROC metrics such as specificity, recall, accuracy, precision, and F-measure using different predictive models as shown in Figure 2. From Figure 2, we can see the factor analysis based XGBoost (FA-XGB)-based predictive model achieves much better performance in comparison with XGBoost and ensemble-based predictive model. Figure 5 shows the ROC performance of different ML-based student performance prediction models.

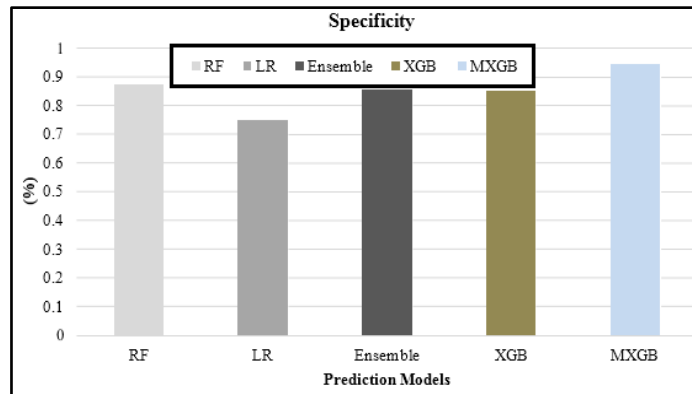


Figure 3. Specificity performance of different ML algorithms for predicting student performance

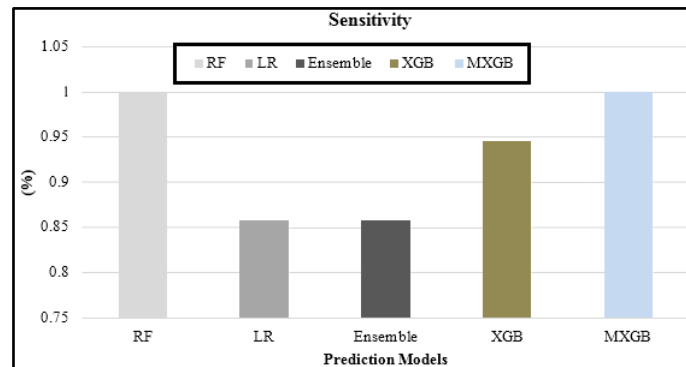


Figure 4. Sensitivity performance of different ML algorithms for predicting student performance

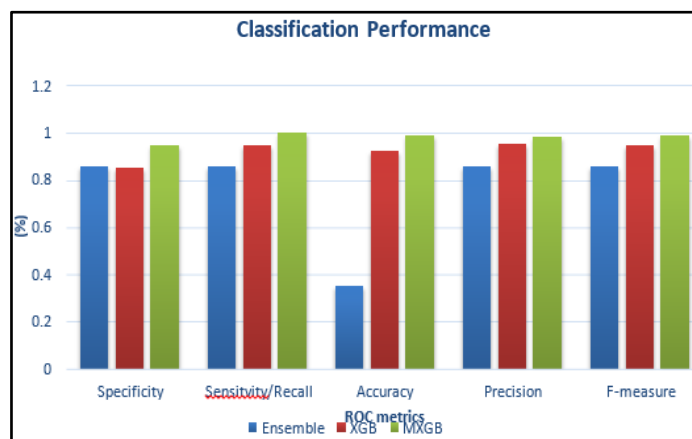


Figure 5. ROC performance of different ML-based student performance prediction models

### 3.2. Feature importance performance

Figure 3 shows a graphical representation of the feature importance parameter obtained using XGBoost and FA-XGB-based predictive model. From Figure 3, we can see that FA-XGB gives higher importance to features in comparison with XGBoost. Further, the FA-XGB-based predictive model gives importance in the following order Kolmogorov-Smirnov (KS), weight (WT), majorization-minimization (MM), moving window (MW), machine learning-based checker (MLC), machine reading comprehension (MRC), and moving window classifier (MWC). On the other side, the XGB-based predictive model gives importance in the following order WT, KS, MW, MM, MRC, MLC, and MWC. Further, it is noticed in both cases MWC is given very less importance. Figure 6 shows how selecting the right feature aid in improving the overall classification accuracy of the proposed FA-XGB-based predictive model.

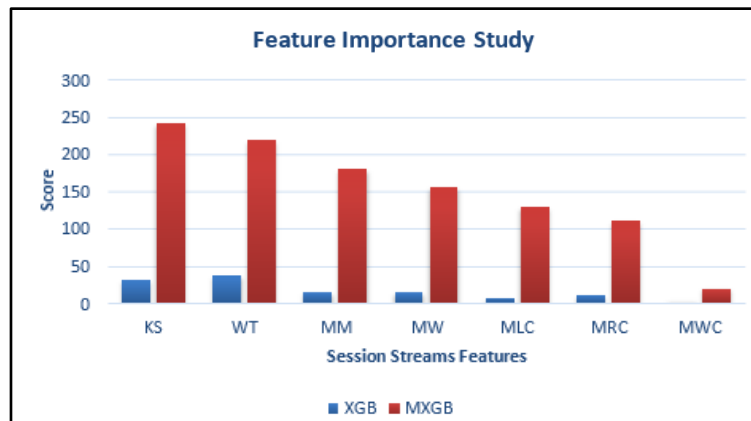


Figure 6. Feature ranking score graphical representation

### 3.3. Student performance prediction for a different session

Here the performance is validated considering different ROC metrics such as specificity, recall, accuracy, precision, and F-measure for different sessions such as session 2, session 3, session 4, session 5, and session 6 using a different predictive model such as XGBoost and FA-XGB as shown in Figures 4 to 8, respectively. Figure 7 shows the accuracy performance using ML-based student performance prediction model for different sessions. From Figures 4 to 8 we can see the FA-XGB-based predictive model achieves much better ROC performance in comparison with the XGBoost-based predictive model. Figures 9 to 11 show the specificity, precision, and F-measure performance using an ML-based student performance prediction model for different sessions.

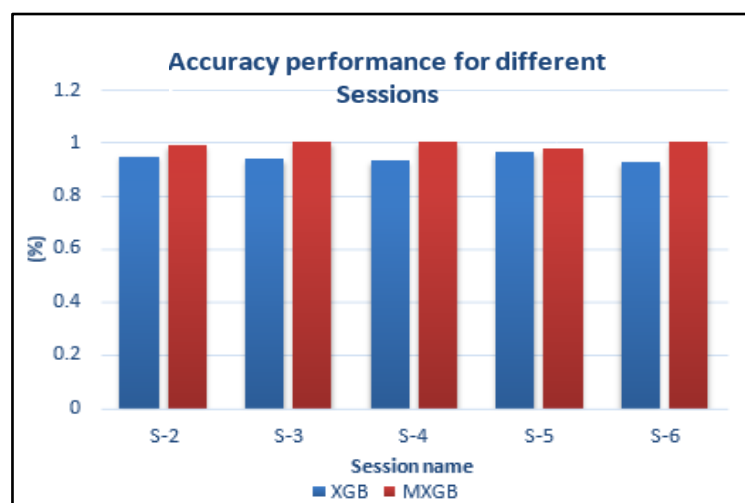


Figure 7. Accuracy performance using ML-based student performance prediction model for different sessions



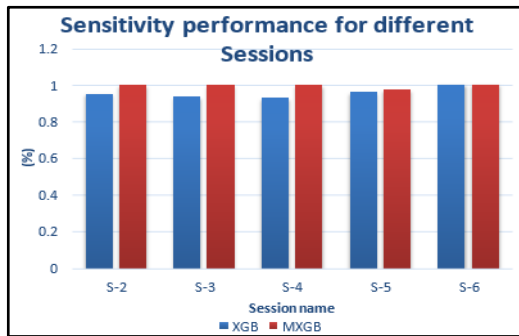


Figure 8. Sensitivity performance using ML-based student performance prediction model for different sessions

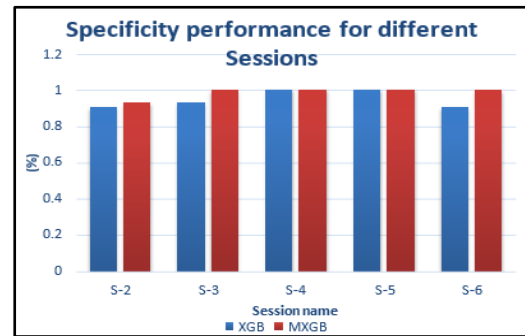


Figure 9. Specificity performance using ML-based student performance prediction model for different sessions

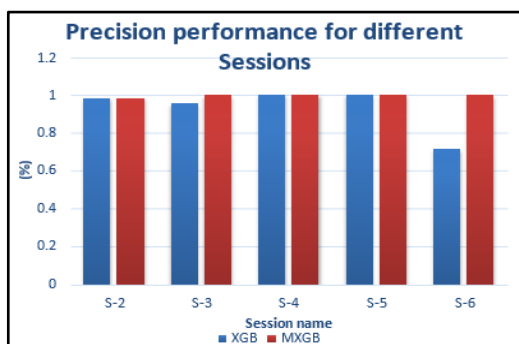


Figure 10. Precision performance using ML-based student performance prediction model for different sessions

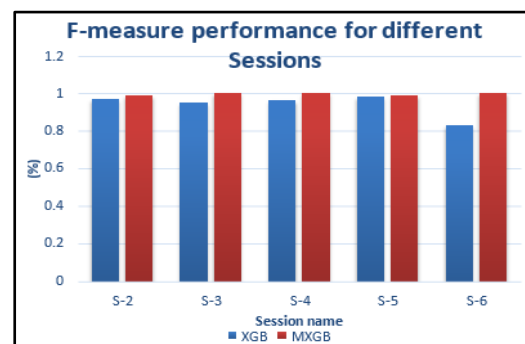


Figure 11. F-measure performance using ML-based student performance prediction model for different sessions

### 3.4. Feature ranking importance

The graphical representation of the feature ranking score of the XGBoost-based and MXGB-based student performance prediction model for different sessions is shown in Figures 12 to 16. Figure 11 shows the graphical representation of the feature ranking score attained using XGBoost-based and MXGB-based student performance prediction model for session 2. From the result it can be stated that XGBoost-based gives a higher score for MW and a lesser score for MRC; On the other side, MXGB-based gives a higher score to WT and a lesser score for MRC. Figure 12 shows the graphical representation of the feature ranking score attained using the XGBoost-based and MXGB-based student performance prediction model for session 3. From the result, it can be stated both XGB-based and MXGB-based give higher scores for MM and lesser scores for MWC; however, the MXGB-based model gives much higher feature importance in comparison with XGBoost-based student performance predictions. Figure 13 shows the graphical representation of the feature ranking score attained using the XGBoost-based and MXGB-based student performance prediction model for session 4. From the result, it can be stated that XGBoost-based gives a higher score for MW and a lesser score for KS, MWC, and MRC; On the other side, MXGB-based gives a higher score to MM and a lesser score to MWC. Figure 14 shows the graphical representation of the feature ranking score attained using the XGBoost-based and MXGB-based student performance prediction model for session 5. From the result, it can be stated that XGBoost-based gives a higher score for KS and WT and a lesser score for MW, MM, and MWC; On the other side, MXGB-based gives a higher score to KS and a lesser score to MWC. Figure 15 shows the graphical representation of the feature ranking score attained using the XGBoost-based and MXGB-based student performance prediction model for session 6. From the result it can be stated that XGBoost-based gives a higher score for KS and a lesser score for MLC and MWC; On the other side, MXGB-based gives a higher score to KS and a lesser score to MWC. The graphical representation from Figures 11 to 15 shows the MXGB-based gives higher importance to features in comparison with the XGBoost-based student performance prediction model. Thus, aiding the MXGB-based student performance prediction model to achieve higher accuracy in comparison with ensemble-based and XGBoost-based student performance prediction models.

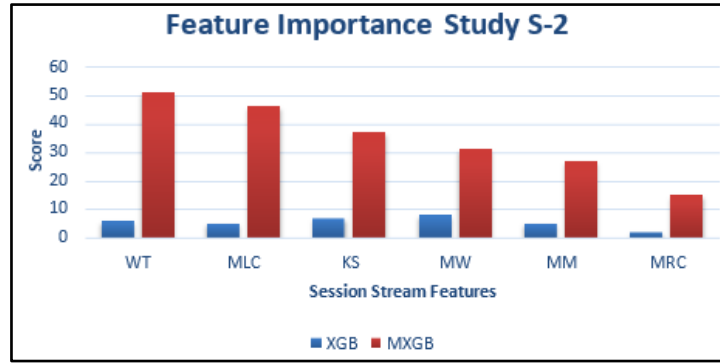


Figure 12. Feature ranking score graphical representation for session 2

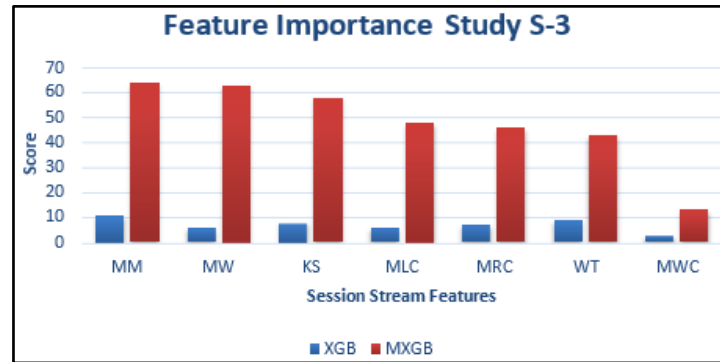


Figure 13. Feature ranking score graphical representation for session 3

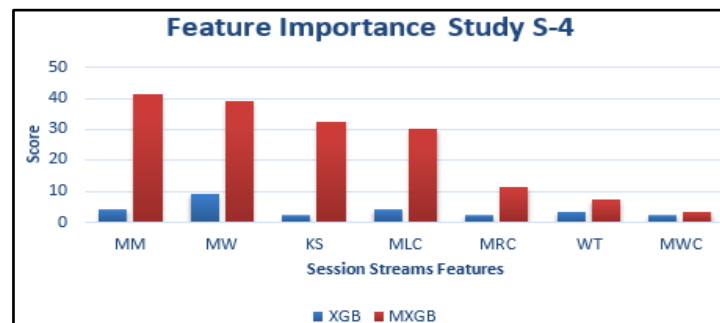


Figure 14. Feature ranking score graphical representation for session 4

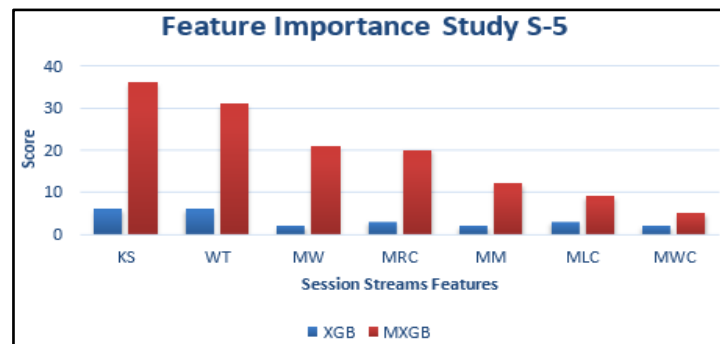


Figure 15. Feature ranking score graphical representation for session 5

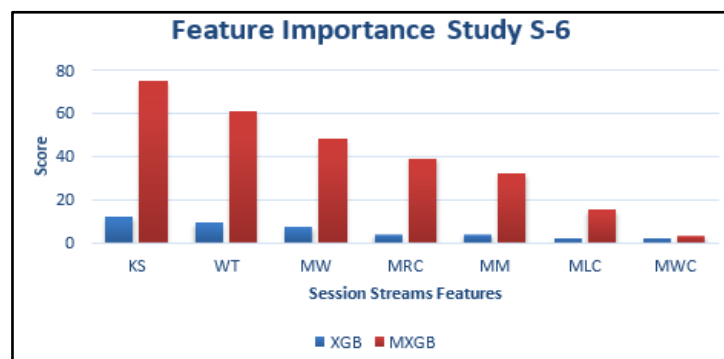


Figure 16. Feature ranking score graphical representation for session 6

#### 4. CONCLUSION

Predicting the performance of a student by analyzing the student session stream is a challenging task. ML algorithms have been used by various existing student performance prediction models to achieve improved prediction outcomes. However, these models tend to achieve higher accuracy to specific student data and when adapted to new data they exhibit poor performance. In addressing such issues, recent work has used an ensemble-based ML model for choosing the best model to perform prediction tasks. However, when data is imbalanced existing ensemble-based models exhibit poor performance. This paper presented an efficient ensemble machine-learning model by modifying XGBoost that works well even when training data is imbalanced. Here an effective cross-validation scheme is presented to identify which feature impacts the accuracy of the prediction model. The cross-validation scheme employs an effective feature ranking mechanism to improve prediction accuracy by optimizing the prediction error. The experiment is conducted using standard student session stream data. The proposed MXGB model significantly improves accuracy, sensitivity, specificity, precision, and F-measure performance in comparison with RF-based, LR-based, ensemble-based, and XGBoost-based student performance prediction models. The performance of the MXGB model will be tested using a more diverse dataset. Alongside this, would consider reducing training errors by considering multi-class classification.




#### REFERENCES

- [1] A. Moubayed, M. Injadat, A. B. Nassif, H. Lutfiyya, and A. Shami, "E-Learning: challenges and research opportunities using machine learning & data analytics," *IEEE Access*, vol. 6, pp. 39117–39138, 2018, doi: 10.1109/ACCESS.2018.2851790.
- [2] F. Essalmi, L. J. B. Ayed, M. Jemni, S. Graf, and Kinshuk, "Generalized metrics for the analysis of E-learning personalization strategies," *Computers in Human Behavior*, vol. 48, pp. 310–322, Jul. 2015, doi: 10.1016/j.chb.2014.12.050.
- [3] J. Yang, J. Ma, and S. K. Howard, "Usage profiling from mobile applications: a case study of online activity for Australian primary schools," *Knowledge-Based Systems*, vol. 191, Mar. 2020, doi: 10.1016/j.knosys.2019.105214.
- [4] A. Wakjira and S. Bhattacharya, "Predicting student engagement in the online learning environment," *International Journal of Web-Based Learning and Teaching Technologies*, vol. 16, no. 6, pp. 1–21, Oct. 2021, doi: 10.4018/IJWLTT.287095.
- [5] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student engagement predictions in an e-learning system and their impact on student course assessment scores," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–21, Oct. 2018, doi: 10.1155/2018/6347186.
- [6] G. Kaur and W. Singh, "Prediction of student performance using weka tool," *Research Cell: An International Journal of Engineering Sciences*, vol. 17, no. January, pp. 2229–6913, 2016.
- [7] Y. Chen, Y. Mao, H. Liang, S. Yu, Y. Wei, and S. Leng, "Data poison detection schemes for distributed machine learning," *IEEE Access*, vol. 8, pp. 7442–7454, 2020, doi: 10.1109/ACCESS.2019.2962525.
- [8] A. J. Stimpson and M. L. Cummings, "Assessing intervention timing in computer-based education using machine learning algorithms," *IEEE Access*, vol. 2, pp. 78–87, 2014, doi: 10.1109/ACCESS.2014.2303071.
- [9] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, 2020, doi: 10.1186/s41239-020-0177-7.
- [10] M. Injadat, F. Salo, A. B. Nassif, A. Essex, and A. Shami, "Bayesian optimization with machine learning algorithms towards anomaly detection," in *2018 IEEE Global Communications Conference (GLOBECOM)*, IEEE, Dec. 2018, pp. 1–6. doi: 10.1109/GLOCOM.2018.8647714.
- [11] L. Yang, A. Moubayed, I. Hamieh, and A. Shami, "Tree-based intelligent intrusion detection system in internet of vehicles," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6. doi: 10.1109/GLOBECOM38437.2019.9013892.
- [12] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "DNS typo-squatting domain detection: a data analytics & machine learning based approach," in *2018 IEEE Global Communications Conference (GLOBECOM)*, IEEE, Dec. 2018, pp. 1–7. doi: 10.1109/GLOCOM.2018.8647679.
- [13] A. Namoun and A. Alshamqiti, "Predicting student performance using data mining and learning analytics techniques: a systematic literature review," *Applied Sciences*, vol. 11, no. 1, Dec. 2020, doi: 10.3390/app11010237.
- [14] S. Ayouni, F. Hajje, M. Maddeh, and S. Al-Otaibi, "A new ML-based approach to enhance student engagement in online environment," *PLOS ONE*, vol. 16, no. 11, Nov. 2021, doi: 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," *IEEE Access*, vol. 9, pp. 69573–69587, 2021, doi: 10.1109/ACCESS.2021.3077663.
- [16] S. S. Khanal, P. W. C. Prasad, A. Alsadoon, and A. Maag, "A systematic review: machine learning based recommendation systems for e-learning," *Education and Information Technologies*, vol. 25, no. 4, pp. 2635–2664, Jul. 2020, doi: 10.1007/s10639-019-10063-9.
- [17] S. Helal *et al.*, "Predicting academic performance by considering student heterogeneity," *Knowledge-Based Systems*, vol. 161, pp. 134–146, Dec. 2018, doi: 10.1016/j.knosys.2018.07.042.
- [18] L. Juhaňák, J. Zounek, and L. Rohlíková, "Using process mining to analyze students' quiz-taking behavior patterns in a learning management system," *Computers in Human Behavior*, vol. 92, pp. 496–506, Mar. 2019, doi: 10.1016/j.chb.2017.12.015.
- [19] Q. Liu *et al.*, "Exploiting cognitive structure for adaptive learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA: ACM, Jul. 2019, pp. 627–635. doi: 10.1145/3292500.3330922.
- [20] F. Wang *et al.*, "Neural cognitive diagnosis for intelligent education systems," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 6153–6161, Apr. 2020, doi: 10.1609/aaai.v34i04.6080.
- [21] B. Kehrwald, "Understanding social presence in text-based online learning environments," *Distance Education*, vol. 29, no. 1, pp. 89–106, May 2008, doi: 10.1080/01587910802004860.
- [22] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowledge-Based Systems*, vol. 200, Jul. 2020, doi: 10.1016/j.knosys.2020.105992.
- [23] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Multi-split optimized bagging ensemble model selection for multi-class educational data mining," *Applied Intelligence*, vol. 50, no. 12, pp. 4506–4528, Dec. 2020, doi: 10.1007/s10489-020-01776-3.
- [24] K. Abe, "Data mining and machine learning applications for educational big data in the university," in *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCCom/CyberSciTech)*, IEEE, Aug. 2019, pp. 350–355. doi: 10.1109/DASC/PiCom/CBDCCom/CyberSciTech.2019.00071.
- [25] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

## BIOGRAPHIES OF AUTHORS



**Shashirekha Hanumanthappa**    currently working as assistant professor in the Department of Computer Science and Engineering, Visvesvaraya Technological University Centre for Post Graduation Studies, Mysuru. She has completed M.Tech. in computer science and engineering from UBDT College of Engineering (Kuvempu University), Davanagere, Karnataka, India in the year 2008. Her field of interest is big data, artificial intelligence, and machine learning. She can be contacted at email: shashirekha\_h2k22@rediffmail.com or shashivtu@gmail.com.



**Dr. Chetana Prakash**    holds Ph.D. in computer science and engineering and she is currently working as professor in the Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere. She has teaching experience of more than 30 years. Her field of interest is speech signal processing, data mining, image processing, fuzzy techniques, IoT, and data analytics. She can be contacted at email: chetana.p.m@gmail.com.