ENHANCING STUDENT PERFORMANCE PREDICTION THROUGH STREAM-BASED ANALYSIS DATASET USING MODIFIED XGBOOST ALGORITHM

Nityashree Nadar

S.I.W.S. N.R. Swamy College of Commerce & Economics and Smt. Thirumalai College of Science, Wadala, Mumbai
India

* Corresponding Author, e-mail: nityashreenadar15@gmail.com

Abstract: In the education domain, predicting the academic performance of students has become an essential task to improve learning outcomes. In this study, we propose a Modified XGBoost (MXGB) model for predicting student performance using stream-based analysis of the dataset. We used a modified version of the XGBoost algorithm using cross-validation, which incorporates stream-based analysis to enhance its performance on real-time data. We preprocessed the dataset and applied feature engineering techniques to extract relevant features for building the model. We trained the MXGB model on the preprocessed dataset and evaluated its performance using various metrics such as accuracy, precision, sensitivity, and F1-score. The results show that our model outperforms the baseline XGBoost model and achieves high accuracy in predicting the student's academic performance. Our model can assist educational institutions in identifying students who are at risk of performing poorly and providing them with timely intervention to improve their academic outcomes.

Keywords: Education Domain, XGBoost, Cross Validation, Feature importance, Machine Learning.

1. INTRODUCTION

In the field of education, predicting student performance has become a critical task to help educational institutions identify students who may need additional support and intervention to improve their academic outcomes. Machine learning algorithms have shown great potential in predicting student performance by analyzing various student attributes such as demographics, academic scores, and extracurricular activities. Moreover, the use of machine learning algorithms for predicting student performance has several advantages over traditional methods such

as an expert judgment or statistical models. One of the greatest difficulties of the 21st century is developing methods to offer knowledge to students on an eLearning platform that is customized to their particular interests and habits. The importance of using student profiles to inform adaptive personalization strategies has been emphasized. Furthermore, various models for predicting student's future performance have been developed using data mining and machine-learning. Data mining is often used for the eLearning platform's session streaming data to extract usable knowledge for the benefit of the students. Also, it enhances the effectiveness of decision-making by learning from past actions. Several sectors, like finance, information security, and education, stand to benefit greatly from the application of machine-learning and data-mining techniques. Education-Data Mining (EDM) is a new field of study that aims to optimize student's types of learners, gain insight into their behaviors, and boost their academic outcomes. Several types of information, such as administrative logs, session streaming activities, and student grades, make up the EDM data. In this study, we focus on predicting student performance using a Modified XGBoost method. The proposed model has been built on the existing XGBoost algorithm. Also, for enhancing the performance for the student performance prediction, a cross-validation technique has been proposed. Further, the significance of the research work is as follows.

- The presented approach for predicting student performance makes use of an effective ensemble-based prediction approach implemented in MXGB, which performs effectively even in the presence of data imbalance.
- The MXGB has a more refined cross-validation technique for investigating what factors influence the performance of a prediction model for students.
- In contrast with the state-of-the-art student performance prediction method, the suggested approach achieves improved ROC performance.

2. LITERATURE SURVEY

In [1], a review of how student performance prediction is done using the EDM data has been done. They have explained various data mining as well as machine learning algorithms and how they can be applied to EDM data. They have performed various experiments using different machine learning algorithms using an education dataset. They have addressed the given various challenges and issues which occur during the prediction of student performance. In [2], they have proposed a machine learning model for the prediction of the grades for the final examination of the student using a Turkey student dataset. They have compared their model with various existing machine learning algorithms. The proposed model attained an accuracy of 75% when compared with the existing models. Further, in [3], they collected the student's data from the BMS College of Engineering. They have used the WEKA software to analyze the EDM data and predict the performance of students using various machine learning algorithms. The results show that the Random Forest attained the highest accuracy when compared with the other existing algorithms. In

[4], they predicted the performance of the student using online session data. This dataset contains various sessions such as mouse clicks, keystrokes on the mouse, time spent, etc. They have evaluated their model using 5 machine learning algorithms. In all the machine learning algorithms they have used cross validation techniques and have split the data in the ratio of 80:20 for the training as well as testing. The results show that by using random forest with cross validation, it attained an accuracy of 97.4%. In [5], they proposed an ensemble classifier for the prediction of the performance of the student using the EDM dataset. The dataset comprises various intercommunication among the students. They have compared their model with various existing machine learning algorithms and the results show that the proposed Ensemble based Decision Tree method has attained an accuracy of 90.4% to 91.4%. The results show that it has attained the highest accuracy when compared with the previously existing methods. In [6] and [7], they have given an EDM dataset which they have collected from various sources and different e-learning systems. In [6] and [7], they have proposed various machine learning methods for the prediction of the student's performance. The results have been evaluated by considering the ROC metrics. The dataset they have provided has class imbalance issues. In both these models, they have not addressed the class imbalance issue and have not selected the important features. Due to this, both models fail to attain higher accuracy.

3. MACHINE LEARNING MODEL FOR EDUCATION DATA MINING OF STUDENT SESSION STREAMS

In this section, we present an enhanced machine learning model for the streams of the Education Data Mining (EDM) called Modified XGB (MXGB). The MXGB is built on the XGBoost technique and is enhanced by using a feature selection technique. Consider an EDM dataset, which has various streams in it. The EDM dataset can be denoted using the following equation:

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

In Equation (1), the row size of the dataset is defined by using $j = 1,2,3, \dots, m$, by considering $b_j \in \{-1,1\}$ which is used for defining the output of the j^{th} row. The a_j is used for representing the $n - dimension$ vector for the self-determination of the features experimentally in the j^{th} row. Moreover, the dataset for the EDM has various features and various values for it. Moreover, the EDM dataset has very less m rows. Hence, to study and design the student-performance prediction technique $\hat{G}$, for the estimation and prediction of the value of G, the following equation has been defined:

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

The proposed architecture of the proposed technique has been given in Figure 1. In this work, the feature-selection technique has been utilized during the training of the XGBoost model which will enhance the performance of the proposed model for the prediction of student performance.

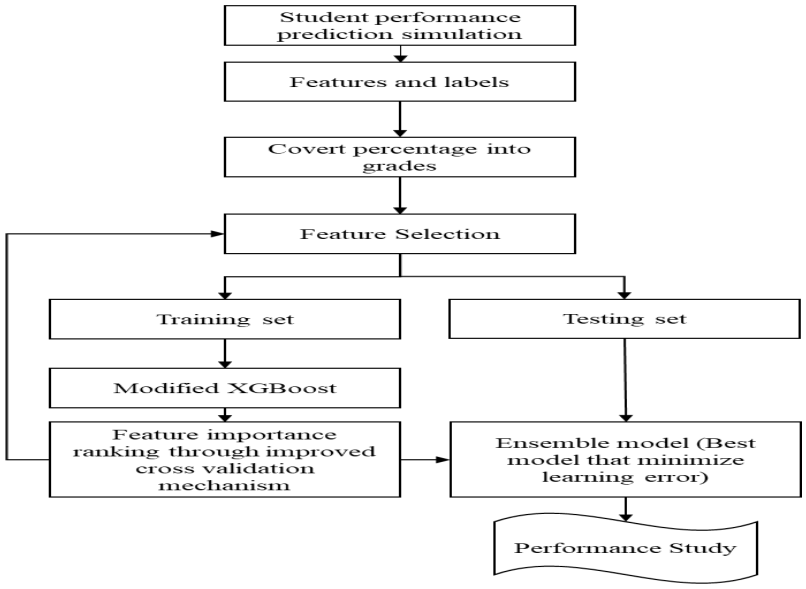

Figure 1. Proposed architecture for the MXGB model for the prediction of student performance using the EDM streams.

3.1. XGBoost Algorithm for Prediction of the student performance

In this section, the XGBoost algorithm has been discussed. The XGBoost method is very much an improved variant of said Gradient Boosting algorithm [8], [9] which is used to generate a strong classifier by combining multiple lesser classifiers to achieve superior classification results. First, in this work, we consider the student streaming session dataset represented as $E = \{(y_j, z_j); j = 1 \dots o, y_j \in \mathcal{S}^n, z_j \in \mathcal{S}\}$, having the dataset samples o with different features n. Consider $\hat{z}_j$ which is used for defining the final output of the XGBoost technique which is represented using the following equation:

$$\hat{z}_j = \sum_{l-1}^{L} g_l(y_j), \;\; g_l \in G \tag{3}$$

In Equation (3), g_l is used for representing the specific tree of regression, $g_l(y_j)$ is used for representing the final output of the XGBoost technique which is given using the j^{th} sample and l^{th} tree. Further the specific tree of regression g_l can be learned utilizing the reduction of the objective function which is defined using the following equation:

$$\mathcal{O} = \sum_{j=1}^{o} m(z_j, \hat{z}_j) + \sum_{l=1}^{L} \beta(g_l) \tag{4}$$

In Equation (4), the m is used for representing the training-loss-function which is helpful to measure the variance between the z_j and $\hat{z}_j$, actual value, and predicted

final output, respectively. To solve the problem of over-fitting in Equation (4), the β parameter has been used. The β parameter will penalize the complexity for the predicted final output using the following equation:

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

The μ and δ in Equation (5) are used for representing the regularization parameter. U is used for representing the size of the leaf, and x is used for representing the score of the various leave of the specific tree of regression. For the construction of an ensemble tree, the summation technique has been utilized. Consider $\hat{z}_j^{(u)}$ represents the final predicted output utilizing the u^{th} iteration as well as j^{th} sample. Further to minimize the objective function, a function g_u is added and then the objective function is defined as the given below equation:

$$\mathcal{O}^{(u)} = \sum_{j=1}^{o} m\left(z_j, \hat{z}_j^{(u-1)} + g_u(y_j)\right) + \beta(g_l) \tag{6}$$

Further, for simplifying the above equation, the Taylor-Expansion has been performed on the above equation. By applying the Taylor-Expansion twice, the following equation is obtained:

$$\mathcal{O}^{(u)} = \sum_{j=1}^{o}\left[h_j g_j(y_j) + \frac{1}{2} i_j g_u(y_j)^2\right] + \beta(g_l) \tag{7}$$

where h_j and i_j are variables that determines the first order-gradient relative to m as shown below

$$h_j = \partial_{\hat{z}_j^{(u-1)}} m\left(z_j, \hat{z}_j^{(u-1)}\right) \tag{8}$$

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

Hence, from the above Equation, Equation (8), and Equation (9), the objective-function is defined using the given equation:

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

The above equation is further simplified as the given below equation:

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

In the above equation, the J_k is used for representing the sample set of the leaves of the tree k which is defined using the given equation:

$$J_k = \{j | r(y_j = k)\} \tag{12}$$

In the above equation, the r is used for defining the tree size which is always fixed, further, the x_k^* the optimal weight of the leaves j is attained using the given below equation:

$$x_k^* = \frac{H_k}{I_k + \mu} \tag{13}$$

Further, the optimized size of the tree for the XGBoost technique is attained using the given below equation:

$$\mathcal{O}^* = \frac{1}{2}\sum_{k-1}^{U} \frac{H_k^2}{I_k + \mu} + \delta U \tag{14}$$

In the above equation, H_k and I_k are defined using the given below equations:

$$H_k = \sum_{j \in J_k} h_j \tag{15}$$

$$I_k = \sum_{j \in J_k} i_j \tag{16}$$

In Equation (14), the $\mathcal{O}^*$ is used for representing the optimized size of the tree r. Small the value of r, the more optimized is the tree. Although the XGBoost technique is effective in achieving highly accurate results, improper selection of features in an unknown context and whenever data is uneven (imbalanced) can lead to a drop in accuracy. Hence, to resolve this given issue, an efficient selection features method during the dataset training is given in the below section.

3.2. Prediction Algorithm for the Modified XGBoost

This work improves upon the prediction technique of the industry-standard XGB by modifying the method for selecting features. Optimal cross-validation, characterized by low validating error, enhances the selection of features technique. When developing a predicting approach, the K-fold cross-validation method is employed to provide the best possible results. In this method, the dataset is randomly split among K subsets of similar size. Furthermore, we utilize $K-1$ to build the student prediction approach, as well as utilize the rest of the dataset to minimize the approach prediction-error. Finally, the error of the cross-validation is optimized using the average of the prediction errors for each possible pair of K. Next, the features with the highest weight are prioritized, as well as the student prediction approach only with the lowest cross-validation error is selected, all from a grid l of suitable outcomes. To choose features efficiently, the suggested cross-validation approach splits the process into two parts. The first steps involve extracting key features from larger feature datasets. The features selected in the first step are then used in the second part to develop a reliable approach for predicting student future performance using the streams. The standard method for calculating the error associated with cross-validation using a single fold is given using the following equation:

$$CV(\sigma) = \frac{1}{M}\sum_{k=1}^{K} \sum_{j \in G_{-k}} P\left(b_j, \hat{g}_\sigma^{-k(j)}(y_j, \sigma)\right) \tag{17}$$

Nevertheless, the aforementioned equation does not specify which factor affects the prediction approach accuracy. An important aspect of this work is modelling how efficient cross-validation may be combined with only a feature-selection process that has a significant effect on the accuracy of the predictions it provides. The modelling of the efficient cross-validation is given using the following equation:

$$CV(\sigma) = \frac{1}{SM}\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{j\in G_{-k}} P\left(b_j, \hat{g}_{\sigma}^{-k(j)}\left(y_j, \sigma\right)\right) \tag{18}$$

Further, for the selection of the ideal features represented using $\hat{\sigma}$, it is defined using the following equation:

$$\hat{\sigma} = \underset{\sigma\in\{\sigma_1,\ldots,\sigma_l\}}{arg\,min}\ CV_s(\sigma) \tag{19}$$

Further, in Equation (18), The parameter M indicates the maximum size of the training dataset, $P(\cdot)$ is used for defining the loss function, $\hat{g}_{\sigma}^{-k(j)}(\cdot)$ is used for defining the function which evaluates all the coefficients. To build the most accurate approach for predicting future student performance, we iteratively apply Equation (18). In other words, after the first part, where the learning error is minimized, a variable is sent towards the second part, where the prediction approach understanding, and incorporation of the feature-importance-characteristic is refined. The method of optimizing to achieve an efficient feature is accomplished through the reduction of the objective function using a gradient descent technique. This is how the optimizing technique works. The pseduo-code for the proposed technique has been given below.

Pseudo code for the student performance prediction

Step 1 Load the dataset E
Step 2 Define the final output $\hat{z}_j$
Step 3 Split the dataset
Step 4 Reduce the objective function using $\mathcal{O}$
Step 5 Solve the problem of overfitting of the objective function using $\beta(g_l)$
Step 6 Construct the tree for the dataset E
Step 7 Simplify the tree using the Taylor-Expansion
Step 8 Construct the tree again
Step 9 Train the model
Step 10 Use the efficient cross-validation technique to select the best features for reducing the error
Step 11 Train the model
Step 12 Evaluate the model
Step 13 Calculate the performance of the proposed model

As a result, when compared to state-of-the-art ML-based student performance prediction methods, the suggested MXGB-based student performance prediction approach achieves significantly better prediction performance which has been discussed in the next section.

4. RESULTS AND DISCUSSION

Here, we compare the proposed MXGB to previous ML-based student prediction methods [6] to see how well each performs at predicting students' future performance using the streams dataset. This study evaluates performance using the eLearning dataset provided in [6]. The dataset was chosen after reviewing a comparative research work [6]. In this work, for the evaluation of the proposed model with the existing model, the K-fold has been considered as 5, i.e., K=5. In this work, eta, min_child_weight, max_depth, and learning rate are the hyperparameters that have been tuned. The Python 3 platform is used to create a machine-learning model that predicts the performance of students. Validating any approach that predicts students' future performance requires the utilization of ROC performance measurements like F-Measure, Specificity, Precision, Sensitivity, and Accuracy. Further, the accuracy is evaluated using the given below equation:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{20}$$

The parameters 'sensitivity', 'specificity', 'precision', and 'F-measure' are evaluated using the given below equations:

$$Sensitivity = \frac{TP}{TP + FN} \tag{21}$$

$$Specificity = \frac{TN}{TN + FP} \tag{22}$$

$$Precision = \frac{TP}{TP + FP} \tag{23}$$

$$F - measure = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity} \tag{24}$$

In Equations (20), (21), (22), (23), and (24), TP is used for defining the True-Positive, FP is used for defining the False-Positive, TN is used for defining the True-Negative and FN is used for defining the False-Negative.

4.1. Predictive model performance evaluation

Here, we compare the accuracy of several machine learning ML-based models for predicting the performance of students. Different types of student performance prediction methods (Random Forest, Logistic Regression, Ensemble, XGBoost, and Modified-XGBoost). The results in terms of specificity are given in Figure 2. The Logistic Regression, XGB, Ensemble, and RF methods attained specificity of 0.75, 0.8502, 0.857, and 0.875 respectively. The proposed MXGB method attained a specificity of 0.946. From the results, it can be seen that the MXGB method attains good specificity in comparison to the existing student performance methods. Further, the results in terms of sensitivity are given in Figure 3. The Logistic Regression, XGB, Ensemble, and RF methods attained specificity of 0.857, 0.857, 0.9449, and 1 respectively. The proposed MXGB method attained a specificity of 1. From the

results, it can be seen that the MXGB method as well as the Random Forest attains good sensitivity in comparison to the existing student performance methods. In Figure 4, the evaluation of classification performance using the ROC metrics for the prediction of the performance of the student using the streaming dataset. In this Figure, three models, Ensemble, XGB, and MXGB have been compared. The performance has been evaluated by using the ROC metric. The results show that the MXGB attains a higher ROC metric when compared with the existing Ensemble and XGB methods.

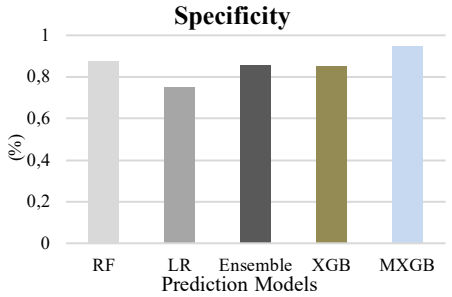

Figure 2. Evaluation of the Specificity Performance for the prediction of the performance of the student using the streaming dataset.

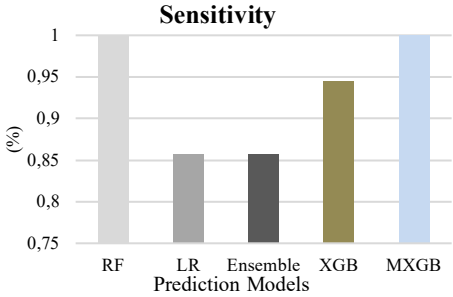

Figure 3. Evaluation of the Sensitivity Performance for the prediction of the performance of the student using the streaming dataset.

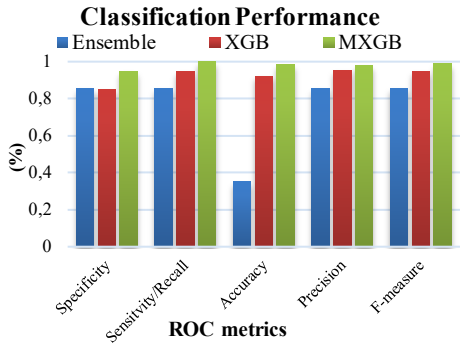

Figure 4. Evaluation of Classification Performance using the ROC metrics

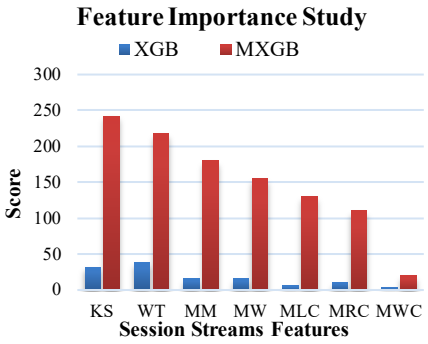

Figure 5. Evaluation of Performance for the Feature Importance.

4.2. Feature importance performance

Here, the study of the performance for the feature importance has been carried out. The performance of the feature importance has been given in Figure 5. The XGB method has been compared with the proposed MXGB method. The features have been taken from the eLearning dataset provided in [6]. The MXGB selects the important features due to which it attains a better score when compared with the existing XGB method.

4.3. Student performance prediction for different sessions

In this section, the performance of the student has been evaluated using the ROC metric. Five sessions, Session 2 (S-2), Session 3 (S-3), Session 4 (S-4), Session 5 (S-5), and Session 6 (S-6) have been evaluated. In Figure 6, the Accuracy for the different sessions by comparing the XGB method and the proposed MXGB method has been given. In Figure 7, the Sensitivity for the different sessions by comparing the XGB method and the proposed MXGB method has been given. In Figure 8, the Specificity for the different sessions by comparing the XGB method and the proposed MXGB method has been given. In Figure 9, the Precision for the different sessions by comparing the XGB method and the proposed MXGB method has been given. In Figure 10, the F1-Score for the different sessions by comparing the XGB method and the proposed MXGB method has been given. It is noted that the proposed MXGB consumes higher time for the student performance prediction due to the cross-validation technique, yet, provides better performance when compared with the existing XGB technique.

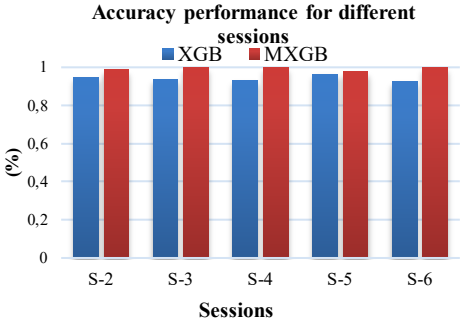

Figure 6. Evaluation of Accuracy Performance for the different sessions using the streaming dataset.

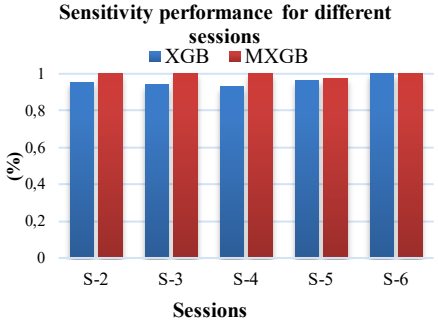

Figure 7. Evaluation of Sensitivity Performance for the different sessions using the streaming dataset.

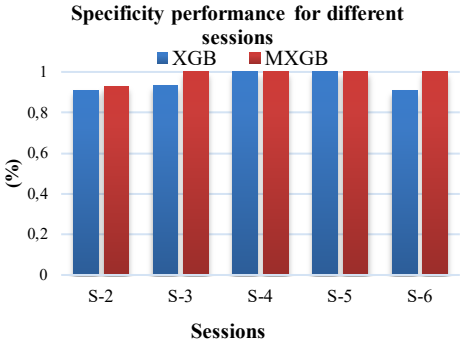

Figure 8. Evaluation of Specificity Performance for the different sessions using the streaming dataset.

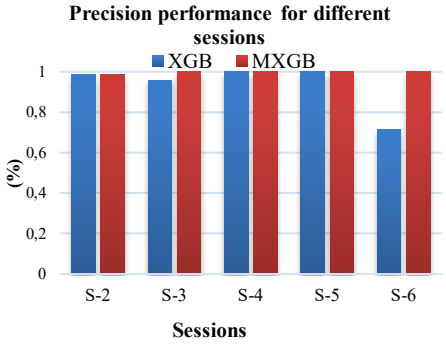

Figure 9. Evaluation of Precision Performance for the different sessions using the streaming dataset.

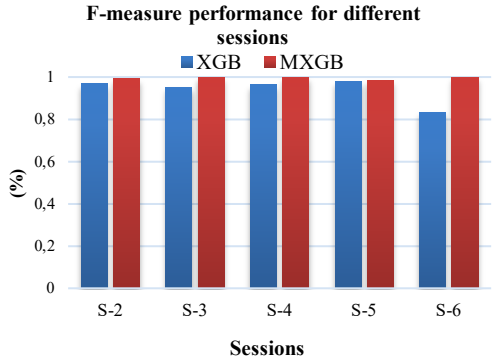

Figure 10. Evaluation of F-Measure Performance for the different sessions using the streaming dataset.

5. CONCLUSION

Predicting the performance of the student by mining their session log data is a complex undertaking. Existing models for predicting students' performance have made use of machine learning algorithms to improve their accuracy. These models are typically more accurate when applied to small subsets of student datasets, but they fail miserably when generalized to larger datasets. As a solution, recent works have implemented ensemble-based ML models to select the optimal approach for the prediction task. Current ensemble-based models, unfortunately, fail miserably when the dataset is unbalanced. This paper modified XGB into an effective ensemble machine-learning method, which performs well despite uneven training examples. In this work, a reliable cross-validation strategy is offered for distinguishing the factors influencing the prediction model's performance. Cross-validation uses an efficient feature rank approach to optimize the prediction-error and boost prediction performance. The experiments have been conducted by using an e-learning dataset that contains various streams. The model has been compared with the previous existing student performance methods by evaluating their performance using the ROC metrics. The results show that the MXGB attain 98.54% which is much higher than the existing methods. For future work, the MXGB method can be tested using different student performance datasets. Also, the training-error for the multi-classification work can be reduced.

REFERENCES

[1] Zhang, Yupei, et al. Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. *Frontiers in Psychology*, vol. 12, 2021, DOI: 10.3389/fpsyg.2021.698490.

[2] Yağcı, Mustafa. Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms. *Smart Learning Environments*, vol. 9, no. 1, 2022, DOI: 10.1186/s40561-022-00192-z.

[3] Arun, D K, et al. Student Academic Performance Prediction Using Educational Data Mining. *2021 International Conference on Computer Communication and Informatics (ICCCI)*, 2021, DOI: 10.1109/iccci50826.2021.9457021.

[4] Brahim, Ghassen Ben. Predicting Student Performance from Online Engagement Activities Using Novel Statistical Features. *Arabian Journal for Science and Engineering*, vol. 47, no. 8, 2022, pp. 10225–10243., DOI: 10.1007/s13369-021-06548-w.

[5] Ragab, Mahmoud, et al. Enhancement of Predicting Students Performance Model Using Ensemble Approaches and Educational Data Mining Techniques. *Wireless Communications and Mobile Computing*, vol. 2021, 2021, pp. 1–9., DOI: 10.1155/2021/6241676.

[6] Injadat, Mohammad Noor, et al. Systematic Ensemble Model Selection Approach for Educational Data Mining. *Knowledge-Based Systems*, vol. 200, 2020, p. 105992., DOI: 10.1016/j.knosys.2020.105992.

[7] Injadat, MohammadNoor, et al. Multi-Split Optimized Bagging Ensemble Model Selection for Multi-Class Educational Data Mining. *Applied Intelligence*, vol. 50, no. 12, 2020, pp. 4506–4528, DOI: 10.1007/s10489-020-01776-3.

[8] Chen, Tianqi, and Carlos Guestrin. XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, DOI: 10.1145/2939672.2939785.

[9] Arati Shahapurkar, S. F. Rodd. Efficient feature aware machine learning model for detecting fraudulent transaction in streaming environment. *International Journal on Information Technologies and Security*, vol.14 , no.3, 2022, pp. 3-14.

Information about the authors:

Dr. Nityashree Nadar is an Assistant Professor in the S.I.W.S. N.R. SWAMY COLLEGE OF COMMERCE & ECONOMICS AND SMT. THIRUMALAI COLLEGE OF SCIENCE. She has been since 2010. Her research interests span machine learning and student analysis. She has published her work in many journals.

Manuscript received on 01 March 2023