JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Student Performance Prediction Model using Modified XG Boost algorithm.

ShashirekhaH[1], Dr.ChetanaPrakash[2]

[1]Asst.Professor, Dept of Computer Science& Engg, VTU Centre for PG Studies, Mysuru.

[2]Professor, Dept of Computer Science& Engg, Bapuji Inst of Engg Tech, Davanagere.

***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 eLearning 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 doesn't 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 MXGB. The MXGB incorporates an effective cross-validation scheme that learns correlation among features more efficiently. The experiment outcome shows the proposed MXGB-a-based student performance prediction model achieves much better prediction accuracy contrary to the state-of-art ensemble-based student performance prediction model.**

***Keyword* – E-Learning, Data Imbalance, Ensemble Algorithm, Feature Importance, Machine Learning**

I . Introduction

The wide usage of the Internet and the growth of information technology have impacted the way academic and industries learns i.e., it is moved from conventional offline mode to online mode namely the eLearning platform [1]. Especially during COVID-19 the pandemic period, all classes have moved to an online model, highlighting the significance of the e-Learning platform. However, significant challenges exist in providing reliable and accurate models to predict student performance [2]. Designing an effective assessment model for understanding student behavior using session streams of the eLearning 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 and machine learning have been used for building student performance prediction models. The data mining has been used for establishing useful insight from student data of the e-Learning platform[6]as shown in Fig.1; alongside, improves decision-making performance [7] [8], [9]. Both machine learning [10], [11], [12] and data mining [13] methodologies are very promising in different fields such as business, and network security including education [14], [15], and [16]. Recently, a new field has emerged namely education data mining (EDM) [17] for enhancing learning style [18], understanding behavior [19], and improving student performance [20]. The EDM data is composed of different information [21] such as administration data, student session stream activity, and student academic performance data. In [22], [23] provided an EDM dataset collected from different databases and e-learning systems. Here different machine learning models and also an ensemble learning mechanism is constructed for predicting student performance during the course. The outcome shows ensemble model outperforms another model in terms of prediction accuracy. However, when data is imbalanced these model fails to establish feature impacting the predictive model; thus, provides poor classification accuracies

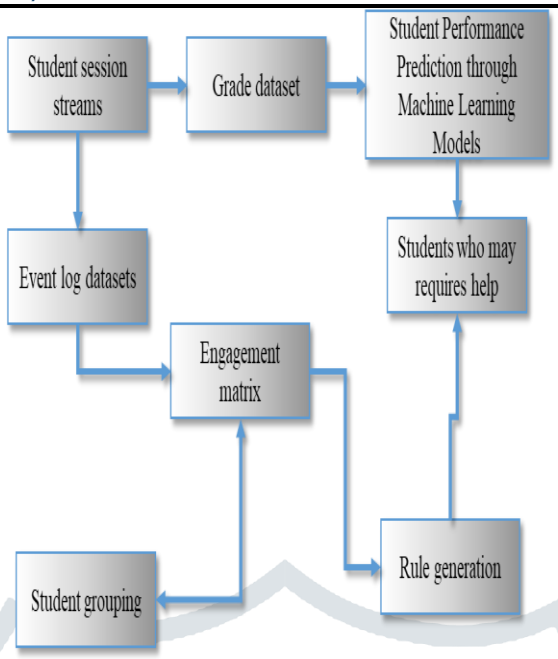

Fig.1.General design of student performance prediction through machine learning models.

II. Literature Survey:

This section surveys 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],	The model focused on analyzing student session stream data of open universities through machine learning models. Good classification accuracy is achieved for the Open University dataset.	However, when employing the model for different student session data these ML-based model performance badly.
Krishna Murthy et al., [8]	Designed Student performance and risk prediction, risk through feedback according to context-based cognitive skill ranks.	The model works only with prior information, of course, is available and when tested under a new environment poor classification accuracy is achieved [7]
Moubay ed et al., [24]	Designed Student engagement level prediction in an e-learning platform employing a K-mean clustering algorithm	The model does not provide a good result when feature sizes are varied considering multiclass classification

Injadat et al., [22],	Designed ensemble learning by combining multiple ML algorithms 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 the training dataset is imbalanced.
Injadat et al.,[23],	Modeled an optimized bagging ensemble learning algorithm for improving the prediction accuracy of student performance.	The model fails to establish the feature impacting performance of the classifier. Poor classification is incurred when data is imbalanced.

2.1 Problem Statement: The objective of this paper is to build an effective student prediction model for predicting student grades during the course through an ensemble-based machine learning model that works well for eLearning data.

2.2 Research Motivation: Existing model construct ensemble learning by combining multiple ML models. However, these models are effective to address the binary classification problem and when put forth an under multi-label classification problem considering data imbalance, these methods exhibit poor accuracy [22],[23]. The aforementioned limitations motivate this research work to develop an improved student performance prediction model through improved ensemble methodology.

2.3 Research significance:

The proposed student performance prediction model employs an efficient ensemble-based predictive model through MXGB which works well even when data is imbalanced.

The MXGB encompasses an improved cross-validation mechanism to study which feature impacts the accuracy of a student prediction model.

The proposed student performance prediction model achieves better ROC performance such as accuracy, sensitivity, specificity, sensitivity, precision, and F-measure comparison with the state-of-art student performance prediction model.

2.4 Proposed Methodology: Proposed methodology presents an effective student performance prediction through an improved ensemble-based ML model. First, the model briefs a detail of the ensemble algorithm namely XG Boost (XGB). Then, discusses the limitation of standard XGB when data is imbalanced. In addressing a modified XGB (MXGB) based student performance prediction model is presented. The MXGB encompasses an improved cross-validation mechanism for establishing features impacting the accuracy of the student performance prediction model. Finally, an ensemble-based ML is constructed for building an effective student performance predictive model.

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

III. Organization of the module:

In section II, Machine learning model for education data mining of student session streams. In section III, 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 MXGB based student performance prediction model over the existing ensemble-based student performance prediction model is discussed

3.1 MACHINE LEARNING MODEL FOR EDUCATION DATA MINING OF STUDENT ACADEMIC PERFORMANCE

This section presents an improved machine learning model namely MXGB for education data mining of student session streams. The MXGB is an improvement of the standard XGB considering an effective feature selection mechanism. The dataset of standard EDM is defined as follows

where j = 1,2,3, …, m, outlines row size considered, $b_j \in \{-1,1\}$ defines j^{th}row output, and $a_j$defines $n-$dimension vector 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 follows

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

In this work by modifying the feature selection process during training XGB through minimization of the objective function an effective student performance prediction model is designed as shown in Fig.2.

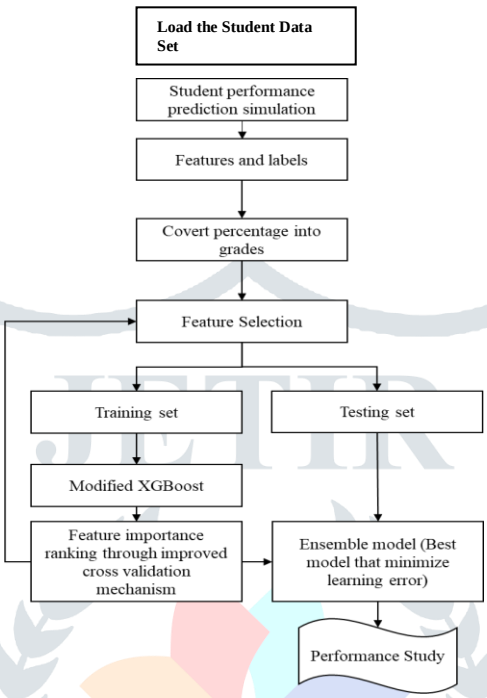

Fig.2.Proposed Machine Learning model for educational data mining of student session streams.

Modified XG Boost Prediction Algorithm.

In this work, the feature selection process of standard XGB 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 Ksubset 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 of K is used for optimizing the cross-validation error. After that, a grid of l appropriate outcomes is obtained for obtaining optimal prediction that minimizes 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 features are selected from feature subsets. In the second phase, the feature chosen from the first phase is utilized for constructing an effective student performance prediction model. The traditional single-fold cross-validation error is constructed as

$$CV(\sigma)=\frac{1}{M}\sum_{k=1}^{K}\sum_{j\varepsilon G-k} P\left(b_{j,}\, \hat{g}_{\sigma}^{-k(j)}(y_{j,}\sigma)\right) \quad (17)$$

However, the above equation doesn't identify which feature impacts the accuracy of a predictive model. In addressing this work an effective cross-validation with effective feature selection with high importance impacting prediction accuracy is modeled as follows

$$CV(\sigma)=\frac{1}{SM}\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{j\varepsilon G-k} P\left(b_{j,}\, \hat{g}_{\sigma}^{-k(j)}(y_{j,}\sigma)\right) \quad (18)$$

In Eq.(18), selecting ideal $\hat{\alpha}$for optimizing the student prediction model is attained as follows

$$\hat{\alpha}=\underset{\sigma\,\varepsilon(\sigma 1\ldots\sigma l)}{\arg\min}\; CV_s(\sigma) \quad (19)$$

In Eq. (18), M defines the size of the training data set considered, $P(\cdot)$defines the loss function, and $\tilde{g}^{k(j)}(\cdot)$ defines a function to compute coefficients. Eq.(18) is executed iteratively for constructing the best student performance prediction model (i.e., its optimization of training error is done in the first phase, and 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 ranking method$r(\cdot)$for constructing a student performance prediction model through the following equation

$$r(a)=\begin{cases}0 & if\ n_j\ is\ not\ selected\\ 1 & if\ n_j\ is\ selected\ as\ optimal\ prediction\ model\ j=1,2,3,....n\end{cases} \quad (20)$$

The feature subset is constructed as follows

$F_s = \{r(n_1), r(n_2),……..r(n_n)\}$ (21)

The ideal feature with a maximum score considering varied K-fold instances is obtained as follows

$Fs_k = \{r(n_1), r(n_2),……..r(n_n)\}$ (22)

Then compute the number of occurrences, the particular feature is selected for K feature subsets having a maximum score, and the final feature subset is obtained as follows

$Fs_{final} = \{f_s(n_1), f_s(n_2),…….. f_s(n_n)\}$ (23)

Where $f_s(*)$ depicts a case when where the n^{th} feature is selected or not and mathematically represented as follows

$$F_s(a)=\begin{cases}0 & if\ q_j\ is\ chosen\ less\ than\ \frac{K}{2}\ times\ j=1,2,3,.....n\\ 1 & if\ q_j\ is\ chosen\ greater\ or\ equal\ to\ \frac{K}{2}\ times\ j=1,2,3,.....n\end{cases} \quad (24)$$

The aforementioned equation is used for the generation of a subset of n' selected features, where n^{th} describe how many time a feature is selected. The EDM 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 several 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.

IV. RESULT AND DISCUSSION

In this section, student performance prediction using the proposed MXGB and other existing ML-based student prediction methods are studied [22]. The eLearning dataset from [22] is used for performance analysis. The selection of the dataset is based on a comparison paper [22]. The model is a machine learning 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 follows

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN} \quad (25)$$

where TP defines true positive, FP defines false positive, TN defines true negative, and FN defines false negative. The sensitivity is computed as follows

$$Sensitivity=\frac{TP}{TP+FN} \quad (26)$$

The Specificity is computed as follows

$$Specificity=\frac{TN}{TN+FP} \quad (27)$$

The Precision is computed as follows

$$Precision=\frac{TP}{TP+FP} \quad (28)$$

The F-Measure is computed as follows

$$F-measure=\frac{2\ x\ Precision\ x\ Sensitivity}{Precision\ x\ Sensitivity} \quad (29)$$

Predictive Model Performance Evaluation:

In this section different ML-based student performance prediction model in terms of specificity and sensitivity is studied.

Figures 3 and 4 show specificity and sensitivity outcomes achieved using different student performance prediction models such as RF-based, LR-based, Ensemble-based, XGB-based and proposed MXGB-based.

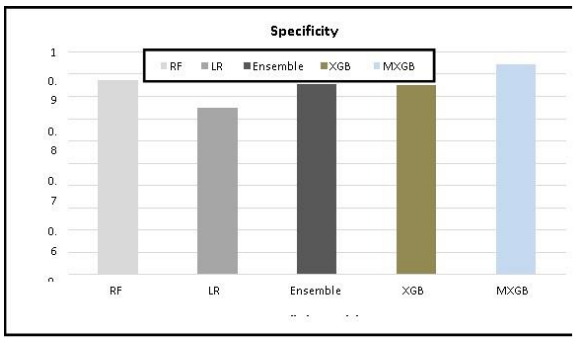

Fig 3. Specificity Performance of different ML Algorithms for predicting student performance

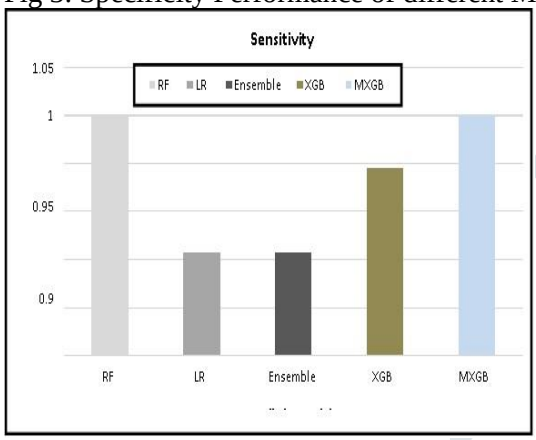

Fig 4. Sensitivity Performance of different ML Algorithms for predicting student performance

Table 1. Comparison of different algorithms with Metrics used.

Metrics	Algorithms Used				
	RF	LR	Ensemble	XGB	MXGB
Specificity	0.875	0.75	0.857	0.85	**0.946**
Sensitivity	1	0.857	0.857	0.9449	**1**

Further, performance is validated considering different ROC metrics such as specificity, recall, accuracy, precision, and F-measure using a different predictive model as shown in Fig.4. From Fig.4 we can see that the MXGB-based predictive model achieves much better performance in comparison with XGB and Ensemble-based predictive model.

Fig 5. ROC Performance of different ML based Algorithms for student performance Prediction Model.

Table 2. Classification Performance Comparison of different algorithms with Metrics used

Metrics	Algorithm Used		
	Ensemble	XGB	MXGB
Specificity	0.85	0.84	**0.964**
Accuracy	0.38	0.94	**0.98**
Sensitivity	0.857	0.95	**1**

Precision	0.84	0.97	**0.99**
F Measure	0.84	0.95	**0.99**

V. Conclusion and Future Work

Predicting the academic performance of a student is a challenging task in an e-Learning Platform. Consider this scenario, various Machine Learning algorithms have been used to predict the academic performance of a student 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 XGB 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 a prediction model. The cross-validation scheme employs an effective feature ranking mechanism to improve prediction accuracy by optimizing the prediction error. The proposed MXGB model significantly improves accuracy, sensitivity, specificity, precision, and F-measure performance in comparison with RF-based, LR-based, ensemble-based, and XGB-based student performance prediction models. Further, future work can be done if the MXGB model would be tested using a more diverse dataset.

Conflicts of Interest Statement

Manuscript title: STUDENT PEFORMANCE PREDICTION MODEL USING MODIFIED XGBOOST ALGORTHIM

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Author names:

1. Shashirekha H
2. Dr. Chetana Prakash

The authors whose names are listed immediately below report the following details of affiliation or involvement in an organization or entity with a financial or non-financial interest in the subject matter or materials discussed in this manuscript. Please specify the nature of the conflict on a separate sheet of paper if the space below is inadequate.

Author names:

No Conflict

REFERENCES

[1]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, Computing Human Behavior .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 the weka tool, Int. J. Eng. Sci. 17(2016).
[7]Dhankhar A., Solanki K., Dalal S., Omdev (2021) Predicting Students Performance Using Educational Data Mining and Learning Analytics: A Systematic Literature Review.In: RajJ.S.,IliyasuA.M., Bestak R., BaigZ.A.(eds) Innovative Data Communication
[8] MD, S., Krishnamoorthy, S. Student performance prediction, risk analysis, and feedback based on context-round 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 the 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] NamounA, AlshanqitiA. 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/app110102
[14] AyouniS, HajjejF, MaddehM, Al-OtaibiS(2021) A new ML-based approach to enhance student engagement in the online environment. PLoSONE16(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.
[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 Proceedings of the 25th ACMSIGKDD 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]B. Kehrwald, Understanding social presence in text-based online learning environments, Dist.Educ. 29 (1)(2008)89–106, http://dx.doi.org/10.1080/01587910802004860.
[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]Chen, Tianqi &Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System.785-794.10.1145/2939672.2939785.

Author Profile

Mrs. Shashirekha H, currently working as Asst. Professor in the Department of Computer Science & Engg, Visvesvaraya Technological University Centre for Post Graduation Studies, Mysuru. She has completed M.Tech in Computer Science & Engg from UBDT College of Engg, Davanagere, Karnataka, India (Kuvempu University) in the year 2008. Her field of interest is Big Data, Artificial Intelligence, Machine Learning.
E-mail: shashivtu@gmail.com

Dr. Chetana Prakash, holds Doctor of Philosophy (Ph.D) in Computer Science and Engineering and she is currently working as Professor in the Department of Computer Science & Engg, Bapuji Institute of Engineering & Technology, Davangere. She has Teaching experience of more than 30 Years. Her field of interest is Speech Signal Processing, Data Mining, Image Processing, Fuzzy Techniques, IoT & Data Analytics. Contact
E-mail: chetana.p.m@gmail.com