European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 08, 2020

Student Risk Identification Model Using Random Forest Algorithm

Susheelamma K H[1], Dr. K M Ravikumar[2]

[1]Assistant Professor, [2]Principal,
[1]Department of ISE, S J C Institute of Technology, Chickballapur - 562101 , Karnataka, India.
[2]S J C Institute of Technology, Chickballapur - 562101, Karnataka, India.

[1]susheela.kh@gmail.com, [2]kmravikumar75@gmail.com

Abstract

The main aim of this work is address the issues and predict students who fail or not complete their online graduation course within stipulated time. Training of the existing machine learning (ML) model is done from existence of data from previous course. This manuscript finds the solution for efficient learning when we don't have learning data from previous years for a particular course (i.e., for the new course introduced which has no history). To address the problem mentioned the proposed work builds a machine learning model which uses data from newly introduced course. For this the proposed model uses newly introduced course data of already submitted task, Hence the model induces imbalanced data issues. For addressing this issue, this work presents a Random Forest (RF) classification algorithm. By the results obtained by experiments conducted we see that a significant outcome is attained by proposed model compared to existing ML models.

Keywords: Virtual Learning Environment, Classification, Imbalanced data, Machine learning, Open Online Courses.

1. INTRODUCTION

Number of student dropout is more in Open Online Courses compared to school levels, higher secondary level and at the graduation level. According to earlier research [1], [2], the students who are unable to complete graduation is 20% in USA, The students who are not completing the studies within a given period in Europe it is 20% to 50% [3]. Statics goes worse if we consider online education, the number of candidates who drop out the graduation course is 78% [4]. The scenario is even worse when we consider students registered under MOOCs, here the student who completes the course successfully is only 5% [5] or 15% as reported in [6]. Various research communities extensively analyzed the problems of students who fail in completing their course [7], [8]. The KDD CUP 2015 competition concentrated on finding out students who will withdraw from online courses.

The initial step is finding out students, who may withdraw or fail in particular course and hence provide them additional study material. In General, instructor/professor carries out the supportive measures based on the information/outcome obtained by forecasting [8]. Forecasting model communicate with students directly by building email messages [9]. The main aim is to aid student to complete the study program by providing necessary study materials, important questions related to course and see that students are active(busy) throughout the course. Most of the open online courses all study materials provided by Virtual Learning Environment (VLE) and all actions done are recorded and backup in VLE. Along with, VLE also keeps the other

information of student like assessment results, and students' demographic information, etc. To erect a forecasting model or predictive model, first we cleanse the raw data to get quality training data later in second step we apply ML model, these models are then used to predict the students who are unable to complete the course within the given period. In general predictive model is build by training the model with legacy data obtained from history of course or the information obtained from the task submitted previously [8]. Further, it is applied to the current situation. However if there are new courses introduced which has no history, there is a need for finding new solution.

With the survey done on Open online courses [11] and on Courses related to Higher Education (HE) it is seen that the more number of dropout happens within a month or first few weeks and during first year's course presentation. The reason for a student to dropout may also be because of course fee payment. Therefore, the aim here is to finding a student who drops the course in middle without completing course within stipulated time because of their irregular learning habit. Rapid student drop out may also happen at later stage of course [10] because course design varies in different universities or in educational institutions.
Further, number of ML [11], [12], [13], [14], and [15] based approaches has been presented to identify student who fail to complete course in a stipulated time. However, these models are not efficient when we need to find out a student at risk of not completing course that have no prior information and with presence of imbalanced data. Further, there are number of approaches [7], [16], and [17] presented solution for solving problem of classification for finding a student who is at-risk of failing when there is presence of imbalanced data. The researches not focused on inactive students they focused only on active students. To overcome the said research challenges, our work aimed to find out a inactive student who is at the risk of not completing course in a stipulated time by designing and building a efficient forecasting model.

Research Contributions are as follows:

- Presenting a Random Forest algorithm for identifying student risk of failure using imbalanced data.
- The Random Forest algorithm is used as a binary classifier.
- The proposed model can attain good accuracy even for forecasting for new courses.
- Experiment outcome shows the proposed model is able to obtain better performance in ROC and F-measure over state-of-art model.

We have organized the next parts of the paper as given below. Part II presents Literature survey pertaining to work for addressing imbalanced data. Section III presents the machine learning model build to identify the at-risk students on time by the use of random forest algorithm. The section IV presents experimental study. The last section describes conclusion and future work.

2. Litterature Survey of Forecasting Model Using Imbalanced Data

This section carry out survey specific to ML based research that worked toward addressing presence of imbalanced data. Generally the modeled ML algorithms are used when training data is balanced in order to learn the objective parameter from the given data. But if we consider the real-world environment, we get imbalanced data (i.e., total number of data in some classes are significantly less than that of other classes). This results performance degradation in state-of-art algorithm [18], [19], [20], [21], [22], [23] and [24] while it is used for identifying the failing student risk probability [9]. In order to attain fine-grained binary classification model we need to fix minority classes weight parameter (i.e., majority class weight is considered as1) to attain fine-grained binary classification model. Further, there are many approaches [7], [16], and [17] to solve the classification problem in the presence of imbalanced data to forecast students unable to clear the course in the given academics. However, they just focus on active students and they have not considered student who has not shown interest in completing the assigned tasks. Further,

models mentioned are inefficient if we have linearly non-separable data hence the model may degrade in classification performance accuracy.

Algorithms that provide probabilistic forecasting are few in number because these forecasting models requires two steps first it orders students based on their likeness to fail, resources constraint is used in the second step. For addressing these issues in next section this work present a model to identify a student at risk using Random forest algorithm to accurately identify student who fail to complete course in a stipulated time.

3. Student Risk Identification Model Using Random Forest Algorithm

Proposed manuscript presents a machine learning model suitable for forecasting build using training data extracted from current course. Completed and pre-submitted task of students are used to identify the students who may fail to submit their assignment. Assumption of behavior pattern of students who submit their tasks in nearby days follows the same behavior pattern of students who already completed their tasks. Same behavior pattern assumption is made even for the students those are completed and submitted their tasks. But the model assumes different behavior pattern for the students those are not completed or submitted their task will be different. Though there are already many ML based classification models are there in this work the classification model is used as a binary classification problem. This model is build based on binary classification algorithm which will classify students to two categories First category will hold all the students who submit their assignments on or before deadline date, Second category will hold all students who are unable to submit their assignments on or before deadline date. Here we consider a day such that the day considered should be k days before the deadline day. Hence we can say the forecasting is done on the deadline day if k=0. Here to do forecasting the students considered are enrolled for the course and not finished the assigned task.

- **System model:**
  Let's take the date when forecasting done and deadline date, the date considered for forecasting is k days before the day of deadline. We establish d deadline date and k forecasting days as template deadline and forecasting days for period [forecasting date; deadline date] to build a proposed forecasting model. The deadline date considered is 6 days from the present day if the model is build to forecast whether a group of students are submitting their assigned task/assessment today or within next 6 days. Training data is obtained from days [task/presentation initilaized+6] = 11 with [present+4; present + 1] = [10; 7 and present day information are inaccessible] as labels of submission. The training and testing data's virtual view of days are considered as follows: day = 0 represents today, negatives values represents the information which is known and positive values represents new/unknown data. For new/unknown data we are not considering the data from previous or older days.

- **Training labeling window and feature selection modeling:**
  According to the description of system model, window sampling for labels grow when we use long-term history i.e., there will be more days for training labels when we consider more number of days before the deadline date. The present day condition considered for training the proposed model is 0 to 5 days before deadline day. Window size of both training and testing labels is $k + 1$ for k days before the deadline. The VLE data is composed of rich data like blogs, videos etc. When algorithm is learned for the given day, the model aligns virtual learning environment features in reverse with respect to time on particular days, i.e., we consider day 0 for present day, day 1 for yesterday, day 2 for day before yesterday and so on. For training the model uses the day when the course initialized as the oldest day. Apart from VLE daily count, the model also uses numerous statistical information related to behavior patterns associated with students in VLE like for how many days/how many hours a student was active in VLE, (i.e., days and time student logged and accessed university website, which are the study materials student accessed, login and logout time etc).

- **Proposed Random Forest based Forecasting model using Imbalanced data:**

Random forest (RF) algorithm based approach is an ensemble based ML approach, this approach builds decision trees (DT) in large number during training process and gives a class that is mean or voted by each individual tree [25]. Random forest was first presented in [25] that added an extra layer of arbitrariness to bagging model. Random Forest algorithm is not used in classification and regression, but also utilized in modeling behavior in feature selection [26], [29]. Bootstrap aggregating (Bagging) technique, is an ensemble method modeled to enhance the accuracy of each forecasting model such as trees [25]. Bootstrapping aids DT to minimize the variance and address issues of over fitting. Let's consider a training dataset $Y = Y_{1,2,\dots,o}$ with response $Z = Z_{1,2,\dots,o}$, bootstrapping will continue L times to choose an arbitrary feature data with replacement of training dataset and fits trees to these feature data's. A treei_l, $(l = 1,2, \dots L)$ will be trained each instance. Post completion of training, the outcome of forecasting model can be obtained by taking the maximum vote from L decision trees or by computing mean of the forecasting from L regression tree. An important things to be noted here is that feature data are chosen with replacement, and the probability with condition that few feature is not chosen post L instance selection which can be described as follows

$$\mathcal{P} = \left(1 - \frac{1}{o}\right) L. \qquad (1)$$

In the bootstrapping process of Random Forest, L is generally equal to o. When o is not high enough, a certain percentage of training feature dataset will not be chosen, and these data is called as out-of-bag feature data. Further, the Random Forest enhances the generic tree growing model, where each candidate will be split in scheme of the tree, then arbitrary subset of the feature data are used rather than selecting certain feature set from all the candidates. Whereas in state-of-art tree based ensemble model, if a few feature set are very strong forecaster for the response, these feature sets will be chosen in most of the schemes. Then, these trees will have high correlation. As a result, weakening the forecasting capabilities. The theoretical details of Random Forest are divided into Random forest convergence theorem and generalization error bound. More detail of RF proof can be obtained from [25], and the proposed model for identifying students at risk using RF is described below.

A Random forest model for forecasting at risk students can consider a set of tree structure classification model $i(y, \mathcal{A}_l)(l = 1,2,3, \dots)$, where the $\mathcal{A}_l$ are identically distributed and independent vectors. Further, it requires an efficiency index to express the accuracy parameter of the Random Forest model, which can be expressed as a boundary condition $\mathcal{B}(.)$as follows

$$\mathcal{B}(y, z) = \vec{V}_l J(i_l(y, \mathcal{A}_l) = z) \max_{k \neq z} \vec{V}_l J(i_l(y, \mathcal{A}_l) = k) \qquad (2)$$

Where $\vec{V}$ depicts an mean parameter, $J(.)$ is the indicator function. First part of this index depicts the mean amount of votes at (y, z) for the right class and the second part depicts to the mean vote for the most classes except the right classes. Accuracy parameter will be genearally higher as the boundry is larger. Then the generalization error $\mathcal{G}''$ is obtained as given below

$$\mathcal{G}'' = \mathcal{P}_{y,z}(\mathcal{B}(y, z) < 0), \qquad (3)$$

Where $\mathcal{P}(.)$ Depicts probability. As the size of trees grows, for almost certainly for all cases, $\mathcal{A}_l$, $\mathcal{G}''$ converges to

$$\mathcal{P}_{y,z}\left(\mathcal{P}_{\mathcal{A}}(i(y,\mathcal{A})=z)-\max_{k\neq z}\mathcal{P}_{\mathcal{A}}(i(y,\mathcal{A})=k)<0\right) \quad (4)$$

The convergence of generalization error depicts that Random Forest model can generate or aid in minimizing the generalization error and it will not overfit the model as more trees are added. The upper limit of $\mathcal{G}''$can be expressed as follows

$$\mathcal{G}'' \leq \frac{\vec{\gamma}(1-t^2)}{t^2} \quad (5)$$

Where $\vec{\gamma}$ is average value of the correlation factor, t is efficiency of an individual tree in Random forest model. It averages with increase in efficiency of individual tree, the Random Forest model will attain higher accuracy of forecasting outcomes.

Further, as discussed above, for increasing the size of individual tree efficiency in the Random Forest model, feature analysis has to be first done to establish the dominant feature data for at risk student. In other word, proper feature selection or selecting right feature for cause of risk have to be identified before performing feature ranking. Then, based on the out-of-bag feature data, all the feature set can be ordered by the forecasting capability with the out-of-bag estimates. In particular, tree-structure classification model in Random Forest that possess most significant features at nodes are expected to have highly associated to the response, so that only significant feature can be chosen from these efficient trees. The proposed Random Forest model identifies at risk students with higher accuracy, the performance attainment is experimentally proven in section

4. Experiment Result And Analysis

Performance evaluation of proposed model over existing models are carried out in this section [24], [27], [28]. Here we have done various experiments for experiment analysis with the publically available dataset [27], [28]. The experiment is carried on windows 10 operating system, 64 bit processor with Intel I-5, RAM of 16 GB and Nvidia CUDA enabled 4GB GPU. Proposed work considers forecasting student who failed to submit their task (first) on time or having chance of failing to submit task. The performance of Proposed Random Forest (PRF) model is compared with existing classification models is done by F-measure and ROC.

- **ROC performance evaluation:**
  ROC performance attained by PRF over the available classification model is presented in this section. Experiments are carried out for different deadline days. We conducted experiments for course A [27], [28] and ROC performance is noted as shown in Fig. 1. PRF got improvement in ROC of 32.52% over existing models. Proposed model attained the efficiency in forecasting for different deadline days scenarios.

- **F-measure performance evaluation:**
  F-measure performance achieved by PRF model over available classification model is presented in this section. Experiments are carried out for course A [27], [28] and performance in terms F-measure is noted as shown in Fig. 2. By the result obtained, we see that the proposed PRF is having improvement in F-measure by 11.53%, 16.31%, 10.86%. Hence for various deadline days scenarios the proposed PRF attained the average F-measure performance improvement of 12.9%.

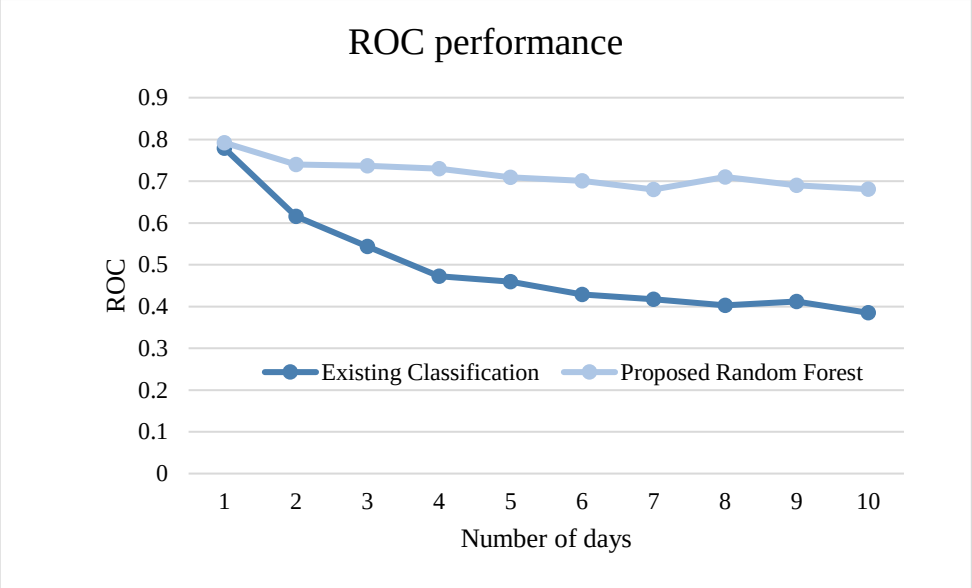

Fig 1: Performance in terms of ROC for varied number of deadline days

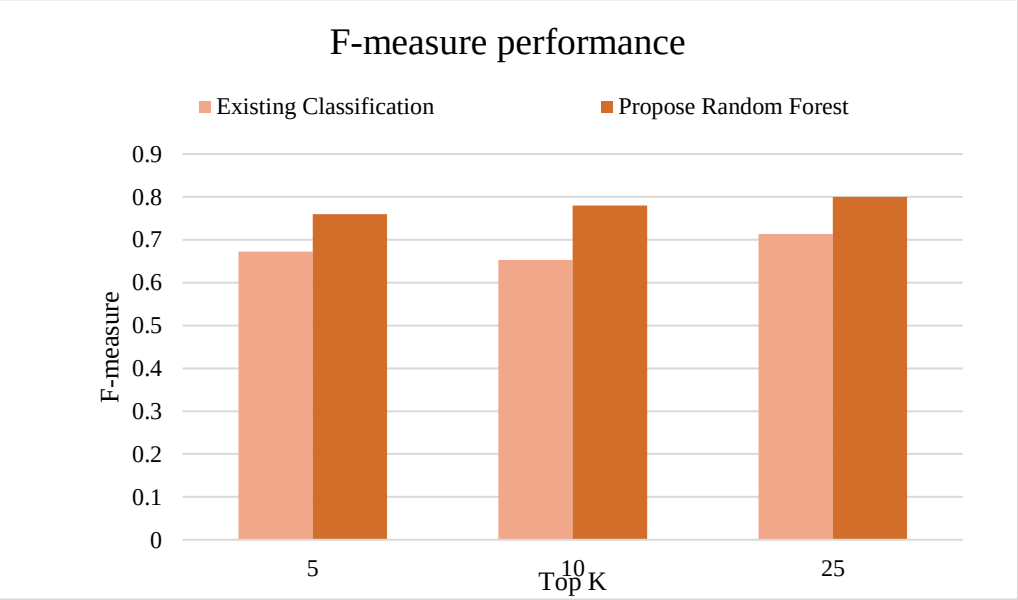

Fig 2: Performance in terms of F-measure for Top K forecasting

5. CONCLUSION

In this work we presented a efficient way of building a machine learning model without using legacy data which detects at risk students who may fail or who may not able to finish the course in a stipulated time. The proposed model learns its pattern by extracting the behavioural pattern of active students who completes and submits the assigned task within time. The proposed work defines the problem as a binary classification with the objective of learning and forecasting by the help of forecasting window. Publicly available OULAD dataset is evaluated in the proposed model. Experimental analysis proved the prediction accuracy of proposed model even for the courses that don't have any past history. It can be seen from overall experiment analysis that in order to forecast a student who is at risk of failing selecting some of the features from VLE is important. The proposed Random Forest based classification model achieves an improvement in F-measure performance by 12.9% over existing models. Further, the proposed model achieves

improvement in ROC performance by 32.52% over the existing model. In future the proposed model would be enhanced to test experimentally different dataset and also the proposed model would be enhanced by the combination of two different algorithms.

REFERENCES

[1] Peter J. Quinn. Drop-out and completion in higher education in europe among students from under-represented groups. Technical report, European Commission, Oct 2013.
[2] H. Vossensteyn, A. Kottmann, B. Jongbloed, and F. Kaiser. Drop-out and completion in higher education in europe executive summary. Technical report, European Commission, 2015.
[3] G. Kena, J. W. X. R. A. Musu-Gillette, Laurenand Robinson, J. Zhang, S. Wilkinson-Flicker, A. Barmer, and E. D. V. Velez. The condition of education 2015. Technical Report 2015-144, NCES, May 2015.
[4] O. Simpson. 22% - can we do better? In The CWP Retention Literature Review, 47, 2010.
[5] K. Jordan. Mooc completion rates: The data. http://www.katyjordan.com/MOOCproject.html, 2015. Accessed: 2017-10-10.
[6] D. Koller, A. Ng, C. Do, and Z. Chen. Retention and intention in massive open online courses: In depth. EDUCAUSE, http://www.educause.edu/ero/article/retention-and-intention-massive-open-online-courses-depth-0, Jun 2013.
[7] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauria, J. R. Regan, and J. D. Baron. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. Journal of Learning Analytics, 1(1), 6-47, 2014.
[8] A. Wol_, Z. Zdrahal, D. Herrmannova, J. Kuzilek, and M. Hlosta. Developing predictive models for early detection of at-risk students on distance learning modules. In Machine Learning and Learning Analytics workshop at LAK14, 24-28 March 2014, Indianapolis, Indiana, USA, 4, 2014.
[9] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart. Predicting mooc dropout over weeks using machine learning methods. In Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, 60-65, 2014.
[10] H. He and E. A. Garcia. Learning from imbalanced data. IEEE Trans. on Knowl. and Data Eng., 21(9):1263-1284, Sep 2009.
[11] C. Taylor, K. Veeramachaneni, and U. O'Reilly. Likely to stop? predicting stopout in massive open online courses. CoRR, abs/1408.3382, 2014.
[12] Wladis, C., Hachey, A. C. & Conway, K., 2014. An investigation of course-level factors as predictors of online STEM course outcomes. Computers & Education, Issue 77, pp. 145-150.
[13] Wolff, A. et al., 2014. Developing predictive models for early detection of at-risk students on distance learning modules. Indianapolis, LAK 2014.
[14] Wang, Rui & Chen, Fanglin & Chen, Zhenyu & Li, Tianxing & Harari, Gabriella & Tignor, Stefanie & Zhou, Xia & Ben-Zeev, Dror & T. Campbell, Andrew. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 10.1145/2632048.2632054, 2014.
[15] Wang, Rui & Chen, Fanglin & Chen, Zhenyu & Li, Tianxing & Harari, Gabriella & Tignor, Stefanie & Zhou, Xia & Ben-Zeev, Dror & T. Campbell, Andrew. StudentLife: Using Smartphones to Assess Mental Health and Academic Performance of College Students. 7-33. 10.1007/978-3-319-51394-2_2, 2017.
[16] J. He, J. Bailey, B. I. Rubinstein, and R. Zhang. Identifying at-risk students in massive open online courses. In AAAI, 1749-1755, 2015.
[17] N. Thai-Nghe, A. Busche, and L. Schmidt-Thieme. Improving academic performance prediction by dealing with class imbalance. In Ninth International Conference on Intelligent Systems Design and Applications, ISDA 2009, Pisa, Italy, November 30-December 2, 2009, 878-883, 2009.
[18] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller,N. Bhanpuri, R. Ghani, and K. L. Addison. A machine learning framework to identify students at risk of adverse academic outcomes. 1909-1918, 2015.

[19] C. Ye and G. Biswas. Early prediction of student dropout and performance in moocs using higher granularity temporal information. Journal of Learning Analytics, 1(3):169-172, 2014.
[20] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison. A machine learning framework to identify students at risk of adverse academic outcomes. 1909-1918, 2015.
[21] E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, and K. L. Addison. Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In LAK '15, 93-102, New York, NY, USA, 2015. ACM.
[22] J. Bainbridge, J. Melitski, A. Zahradnik, E. Lauria a, S. M. Jayaprakash, and J. Baron. Using Learning Analytics to Predict At-Risk Students in Online Graduate Public Affairs and Administration Education. The JPAE Messenger, 21(2):247-262, 2015.
[23] S. Jiang, M. Warschauer, A. E. Williams, D. ODowd, and K. Schenke. Predicting mooc performance with week 1 behavior. In EDM14, 273-275, 2014.
[24] Herrmannova, Drahomira; Hlosta, Martin; Kuzilek, Jakub and Zdrahal, Zdenek (2015). Evaluating Weekly Predictions of At-Risk Students at The Open University: Results and Issues. In: EDEN 2015 Annual Conference Expanding Learning Scenarios: Opening Out the Educational Landscape, 9-12 Jun 2015, Barcelona, Spain.
[25] L. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5-32, 2001.
[26] R. Genuer, J. M. Poggi, et al, “Variable Selection using Random Forests,” Pattern Recognition Letters, vol. 31, no. 14, pp. 2225-1136, 2010.
[27] Kuzilek, M. Hlosta, and Z. Zdrahal. Open university learning analytics dataset. In Data literacy for Learning Analytics workshop at LAK16, 26th April 2016, Edinburgh, UK, 9, 2016.
[28] Kuzilek, Jakub & Hlosta, Martin & Zdráhal, Zdenek, Open University Learning Analytics dataset. Scientific Data. 4. 170171. 10.1038/sdata.2017.171.
[29] Susheelamma K H, Dr. Brahmananda S H, A Survey on Clustering and Feature Selection Algorithm for Quickly Predicting Engineering Students’ Academic Performance, IJSRSET, Volume 4, Issue 1, 2018.