IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SELF-PACED MOOC COURSE STUDENT DROPOUT PREDICTION USING MACHINE LEARNING MODEL

I.Jenita [1],Dr.Jai Ruby[2]

Department of Computer Applications, Sarah Tucker College, Thirunelveli-7.

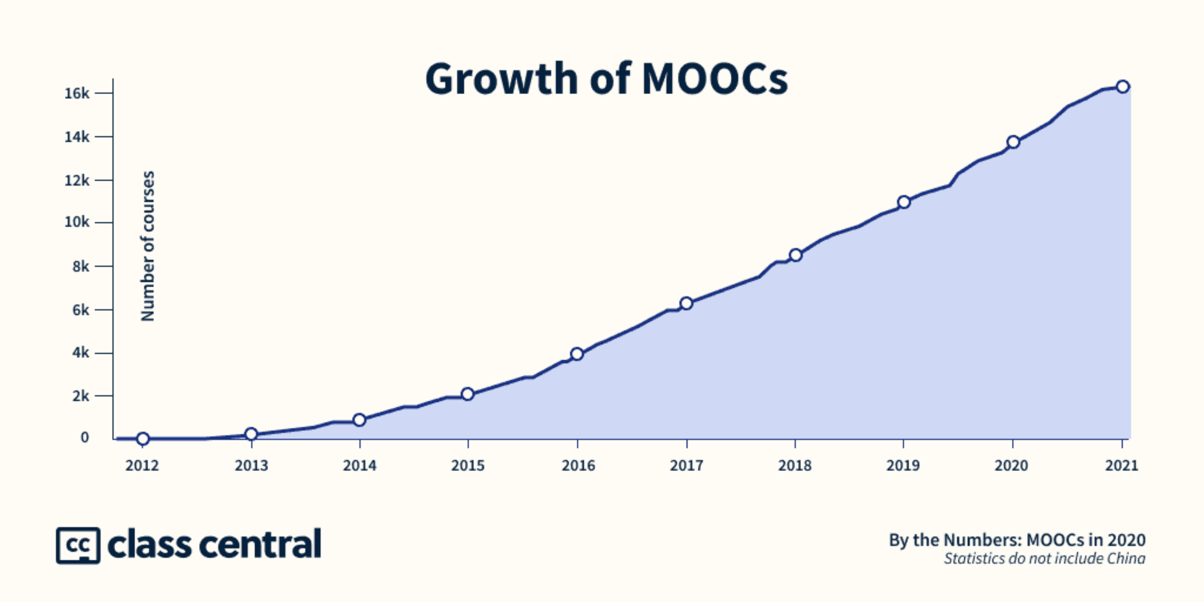
Abstract :

The high rate of student dropout in Massive Open Online Courses (MOOCs) is a serious issue with these courses. An effective MOOC student dropout prediction model can identify the reasons that cause students to drop out and provide insight into how to implement interventions to improve student success. For the prediction of student dropout in MOOC courses, many features and methodologies are available. The data from a self-paced math course, College Algebra and Problem Solving, given on the MOOC platform Open edX in collaboration with Arizona State University (ASU) from 2016 to 2020 are examined in this research. This research proposes a model for predicting student dropout from a MOOC course based on a set of variables derived from the daily learning progress of students.In the prediction, the Gradient Boosting Model technique from Machine Learning (ML) is performed, and validation factors such as accuracy, precision, recall, F1-score, Area Under the Curve (AUC), and Receiver Operating Characteristic (ROC) curve are used. With an accuracy of 87.5 percent, AUC of 94.5 percent, precision of 88 percent, recall of 87.5 percent, and F1-score of 87.5 percent, the model constructed can predict whether students will drop out or continue in the MOOC course on any given day. Shapely values were used to explain the contributing features and interactions for the model prediction

INTRODUCTION:

Dropout prediction in MOOCs is a well-researched problem where we classify which students are likely to persist or drop out of a course. Most research into creating models which can predict outcomes is based on student engagement data.According to a recent Duke University survey MOOC students cited “lack of time/amount of time required” as one of the main reasons for not completing their course. The reading assignments, audio files, videos and homework of online courses take a long time to complete, and courses still follow the same semester format.Whether in the improvement of access worldwide, or the supplementation of programs within existing university communities, it seems MOOCs may well be a leading element in the future of higher education.

MOOCs allow us to keep pace with changes in technology enhanced learning and innovative pedagogies, meeting the strategic aim to design and deliver the best campus-based and online educational experience we can.MOOCs can bring knowledge to students who may not have access otherwise, and be of use to learners who can't afford the costs of higher education. Non-traditional education realised through a MOOCs is a useful form of online learning and can complement traditional university learning.



By the Numbers: MOOCs in 2020
Statistics do not include China

Completion Rates For Online Courses Or MOOCs: What Are They And Do They Matter? A 2018 Columbia University's Teachers College study on edX and Coursera courses shows that MOOC Certificate programs have a completion rate of 15% or less. In the study, learners were asked to voice their suggestions as to why they dropped out the MOOC, and should be done to run these courses more effectively. The results of the analysis of these coded data are given in separate tables below.

Table 1. Reasons for Dropping out MOOC

Personal Reasons	
Obligation to prioritize to other jobs (family, school, work, etc.)	80
Lack of time	56
I was not thinking about completing anyway, I just signed up because I was curious	19
Lack of necessary technology skills	12
Program-related Reasons	
Long course duration	15
Difficult course activities	13
Lack of necessary support from course instructor	12
Others	26

On examination of Table 1, which shows the reasons for dropping out the MOOC, it can be seen that these reasons are gathered under two headings: personal reasons and program-related reasons. Concerning personal reasons, the

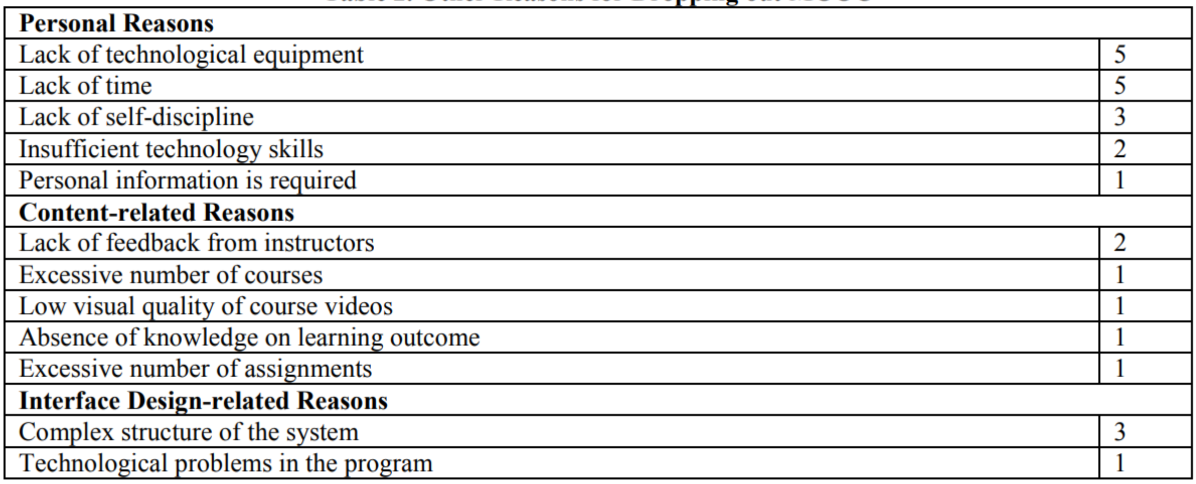
most common reason given is that learners were working, attending an educational institution, or having to spend time with their family (n=80). They also mentioned lack of time as the reason for dropping out (n=56). According to the table, another reason is that some participants (n=19) signed up for the course just because they were curious and did not think whether they would actually complete the course. Finally, it is seen that 12 students dropped the

courses because they did not think they had the technological competence to complete it. Considering the programrelated reasons, it was found that the participants thought that the duration of the courses were too long (n=15),

that course activities were difficult (n=13), or that they did not receive sufficient support from the course instructor (n=12). In addition to these reasons, learners also showed other reasons as to why they had dropped out the courses.

Other reasons are presented in Table 2.

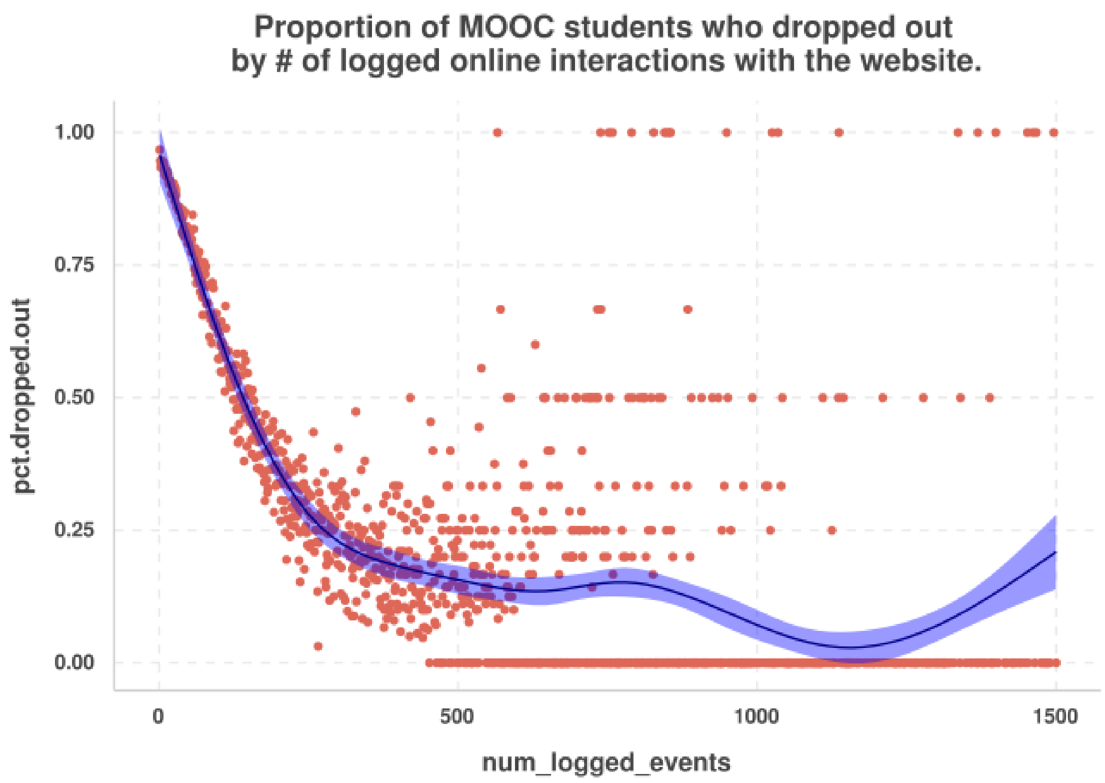
Table 2. Other Reasons for Dropping out MOOC



Personal Reasons	
Lack of technological equipment	5
Lack of time	5
Lack of self-discipline	3
Insufficient technology skills	2
Personal information is required	1
Content-related Reasons	
Lack of feedback from instructors	2
Excessive number of courses	1
Low visual quality of course videos	1
Absence of knowledge on learning outcome	1
Excessive number of assignments	1
Interface Design-related Reasons	
Complex structure of the system	3
Technological problems in the program	1

When Table 2 is examined, it is seen that the opinions of the learners about dropping out the MOOC are grouped under three titles: personal reasons, content-related reasons, and interface design-related reasons. Personal reasons

voiced by learners are as follows: lack of technological equipment (n=5), lack of time (n=5), lack of self-discipline(n=3), lack of technology skills (n=2), and being unwilling to share personal data (n=1). Regarding the reasons arising from content design, learners pointed to a lack of feedback from the course instructors (n=2), the excessive number of courses (n=1), low visual-quality of the course videos (n=1), absence of knowledge on learning outcomes (n=1), and excessive number of assignments given in the courses (n=1). Among the reasons stemming from the interface design, three of the learners found the structure of the system complex, and one learner dropped out the course because the program had technical problems.



LITERATURE SURVEY:

MOOCs are Massive Open Online courses which uses information technologies to boost the learning experience and attract people from the entire world. The learning style of the learners has been changed. The educational domain represents a unique way to deliver the educational knowledge for the learner community with high-quality content throughout the world. The differences between the traditional learning paradigm and digital learning (MOOCs) has brought a new area of research which focus on the online learning.

In [1], a review of literature has been done on students drop out prediction. It also highlight some solutions being used to handle the drop out issues.

In [2], the authors aim to provide a brief and comprehensive review about the challenges that higher education institutions in Macedonia and Kosovo face while coping with the new trends of flexible or blended learning. Moreover, after describing some real cases of MOOC based flipped classroom learning, we also provide some recommendations in order to enhance and enrich learning experience by employing innovative pedagogies.

In [3], the authors have focused on what institutional characteristics contribute to conditions that reduce student dropout risks. By analyzing longitudinal and hierarchical data, this research proposes and tests a multilevel event history model that identifies the major institutional attributes related to student dropout risk in a longitudinal process. Evidence indicates that institutional expenditure on student services is negatively associated with student dropout behavior. Implications of the results for institutional practices and future research are discussed.

In [4], despite the increasing need for STEM skills, to date, the connection between STEM subject choices and their impact on students' educational pathways has not been widely studied. Focusing on the mathematics choice (basic/advanced/no mathematics), a large register dataset that covered students admitted to Finnish universities during 2013–2015 (N = 46,281) was combined with upper-secondary school matriculation examination data (N = 93,955) to find out how this choice influenced the students' university admissions. This large dataset was also examined to establish the current gender distributions in different university degree programs from the perspective of mathematics choices. Further, to find out the students' reasons behind their mathematics choices, a cohort sample of 802 student responses was collected from upper-secondary schools. We also investigated the students' interests in different fields of study to establish any gender differences in them.

In [5], the authors have combined click-stream data and NLP approaches to examine if students' on-line activity and the language they produce in the online discussion forum is predictive of successful class completion. We study this analysis in the context of a subsample of 320 students who completed at least one graded assignment and produced at least 50 words in discussion forums, in a MOOC on educational data mining. The findings indicate that a mix of click-stream data and NLP indices can predict with substantial accuracy (78%) whether students complete the MOOC. This predictive power suggests that student interaction data and language data within a MOOC can help us both to understand student retention in MOOCs and to develop automated signals of student success.

In [6], the methods proposed recently for dropout prediction apply relatively simple machine learning methods like support vector machines and logistic regression, using features that reflect such student activities as lecture video watching and forum activities on a MOOC platform during the study period of a course. Since the features are captured continuously for each student over a period of time, dropout prediction is essentially a time series prediction problem. By regarding dropout prediction as a sequence classification problem, we propose some temporal models for solving it. In particular, based on extensive experiments conducted on two MOOCs offered on Coursera and edX, a recurrent neural network (RNN) model with long short-term memory (LSTM) cells beats the baseline methods as well as our other proposed methods by a large margin.

In [7], the authors have provided an overview of the MOOC student dropout prediction phenomenon where machine learning techniques have been utilized. Furthermore, we highlight some solutions being used to tackle with dropout problem, provide an analysis about the challenges of prediction models, and propose some valuable insights and recommendations that might lead to developing useful and effective machine learning solutions to solve the MOOC dropout problem.

In [8],the authors have investigated MOOC attrition from several different perspectives. Firstly, we review existing literature relating to MOOC dropout rates, bringing together existing findings on completion rates and analyses of several specific courses, which identify factors that correlate to likelihood of dropout. We provide a meta-analysis of the basic figures on overall dropout rates previously collected to identify relationships between course factors and dropout rates. In addition, the literature is reviewed from a qualitative perspective drawing together findings on the reasons for dropout and methods suggested for resolving or reducing the dropout rate. Secondly, using themes emerging from the initial investigation, the authors provide a preliminary analysis of data gathered from a Computing MOOC run by the University of

Warwick, UK and presented using a Moodle platform. Different aspects of students' demographic data are examined to see if relationships to persistence exist. An important feature of this course is that it has been run in two different parallel modes (" traditional " MOOC mode with peer support, and " supported " mode with real time, tutored programming labs).

This allows direct comparison between the dropout figures for the two different modes. Qualitative information from student evaluations is also considered. Finally, we discuss our findings relating MOOC dropout rates, considering what factors are within the control of a MOOC provider and suggesting the most promising avenues for improvement. Our results indicate that many participants who may be classed as dropouts (for example, because they do not complete the necessary components to gain a certificate) are still participating in the course in their own preferred way (either at a slower pace or with selective engagement). This suggests that the structure of " a course " may not be helpful to all participants and supporting different patterns of engagement and presentation of material may be beneficial.

In [9], the authors sought to understand why learners take the courses, specifically *Introduction to Chemistry* or *Data Analysis and Statistical Inference*, and to identify factors both inside and outside of the course setting that impacted engagement and learning. Thirty-six participants in the courses were interviewed, and these students varied in age, experience with the subject matter, and worldwide geographical location. Most of the interviewee statements were neutral in attitude; sentiment analysis of the interview transcripts revealed that 80 percent of the statements that were either extremely positive or negative were found to be positive rather than negative, and this is important because an overall positive climate is known to correlate with higher academic achievement in traditional education settings. When demographic data was added to the sentiment analysis, students who have already earned bachelor's degrees were found to be more positive about the courses than students with either more or less formal education, and this was a highly statistically significant result.

In general, students from America were more critical than students from Africa and Asia, and the sentiments of female participants' comments were generally less positive than those of male participants. An examination of student statements related to motivations revealed that knowledge, work, convenience, and personal interest were the most frequently coded nodes (more generally referred to as “codes”). On the other hand, lack of time was the most prevalently coded barrier for students. Other barriers and challenges cited by the interviewed learners included previous bad classroom experiences with the subject matter, inadequate background, and lack of resources such as money, infrastructure, and internet access. These results are enriched by illustrative quotes from interview transcripts and compared and contrasted with previous findings reported in the literature, and thus this study enhances the field by providing the voices of the learners.

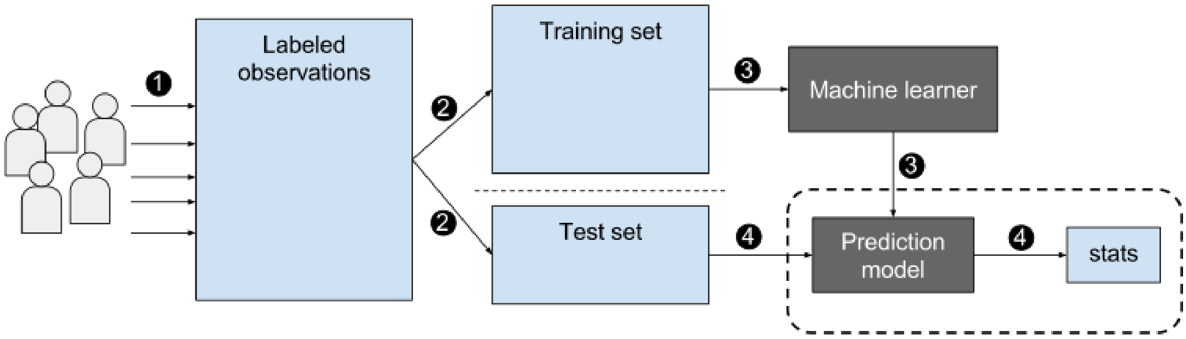
In [10], the report sets out to help decision makers in higher education institutions gain a better understanding of the phenomenon of Massive Online Open Courses (MOOCs) and trends towards greater openness in higher education and to think about the implications for their institutions. The phenomena of MOOCs are described, placing them in the wider context of open education, online learning and the changes that are currently taking place in higher education at a time of globalisation of education and constrained budgets. The report is written from a UK higher education perspective, but is largely informed by the developments in MOOCs from the USA and Canada. A literature review was undertaken focussing on the extensive reporting of MOOCs through scholarly blogs, press releases as well as openly available reports and research papers. This identified current debates about new course provision, the impact of changes in funding and the implications for greater openness in higher education. The theory of disruptive innovation is used to help form the questions of policy and strategy that higher education institutions need to address.

PROPOSED METHODOLOGY :

Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error. i) Gradient Boosting Algorithm is generally used when we want to decrease the Bias error. ii) Gradient Boosting Algorithm can be used in regression as well as classification problems. In regression problems, the cost function is MSE whereas, in classification problems, the cost function is Log-Loss.

GRADIENT BOOSTING ALGORITHM :

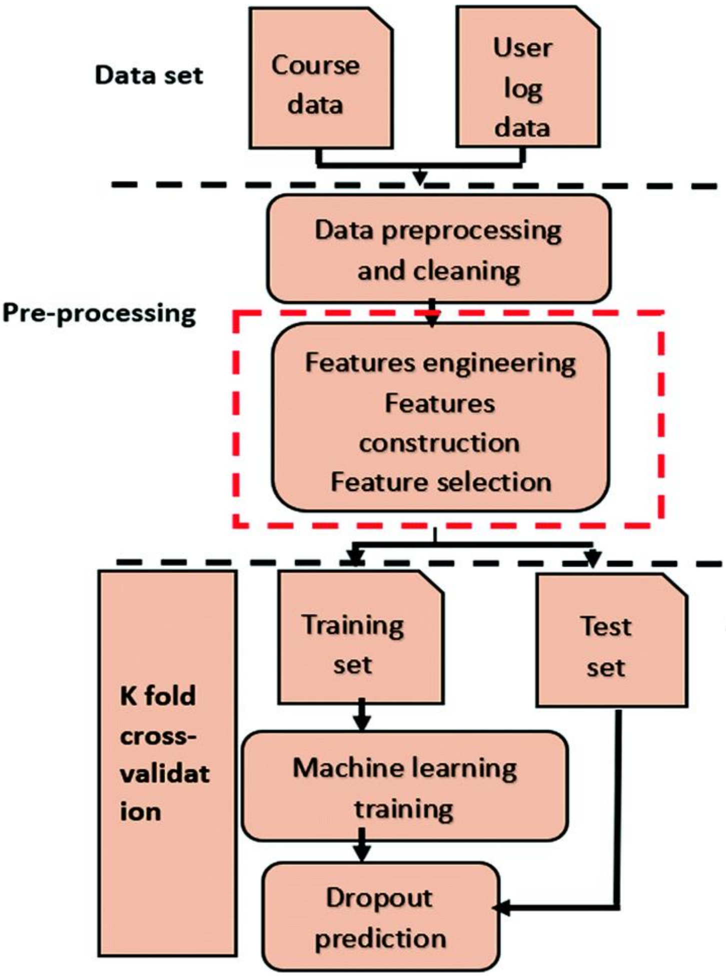
In cases of binary classification (like ours), Gradient Boosting uses a single regression tree to fit on the negative gradient of the binomial deviance loss function. XGBoost, a library for Gradient Boosting, contains a scalable tree boosting algorithm, which is widely used for structured or tabular data, to solve complex classification tasks. Adaboost is another method, performing iterations using a base algorithm. In each interaction, Adaboost uses higher weights for samples misclassified, so that this algorithm focuses more on difficult cases. Random Forest is a method that use a number of decision trees constructed using bootstraping resampling and then applying majority voting or averaging to perform the estimation.



In order to implement a gradient boosting classifier, we'll need to carry out a number of different steps. We'll need to:

- Fit the model
- Tune the model's parameters and Hyperparameters
- Make predictions
- Interpret the results

- Fitting models with Scikit-Learn is fairly easy, as we typically just have to call the fit() command after setting up the model.
- However, tuning the model's hyperparameters requires some active decision making on our part. There are various arguments/hyperparameters we can tune to try and get the best accuracy for the model. One of the ways we can do this is by altering the learning rate of the model. We'll want to check the performance of the model on the training set at different learning rates, and then use the best learning rate to make predictions.
- Predictions can be made in Scikit-Learn very simply by using the predict() function after fitting the classifier. You'll want to predict on the features of the testing dataset, and then compare the predictions to the actual labels. The process of evaluating a classifier typically involves checking the accuracy of the classifier and then tweaking the parameters/hyperparameters of the model until the classifier has an accuracy that the user is satisfied with.



CONCLUSION :

This document gives a summary of research studies that looked into the MOOC student dropout prediction. In several research publications, it also presented an overview of various Machine Learning algorithms used to forecast student dropout. Finally, it discussed the numerous difficulties that exist in predicting student dropout in MOOCs. The near futureWork may be expanded because the scope of research in this subject is so broad, and additional investigations can be conducted.Predictions for the reasons for dropping out. Having a better grasp of why students drop out can help.improve the developer and lectures to improve the course's substance and methodology Data will be the focus of future research a collection of data for assessing and predicting MOOC student outcomes.

REFERENCES :

[1] F. Dalipi, A. Kurti, K. Zdravkova, L. Ahmedi, "Rethinking the conventional learning paradigm towards MOOC based flipped classroom learning‖, 2017 IEEE International Conference on Information Technology Based Higher Education and Trainint (ITHET), July 10-12, 2017, Ohrid, Macedonia.

[2] R. Chen, " Institutional characteristics and college student dropout risks: A multilevel event history analysis". Research in Higher Education, 53(5), 487-505.

[3] L. Ulriksen, L.M. Madsen, H.T. Holmegaard, ―Why do students in stem higher education programmes drop/opt out?–explanations offered from research". In Understanding student participation and choice in science and technology education (pp. 203-217). Springer Netherlands.

[4] S. Crossley, L. Paquette, M. Dascalu, D.S. McNamara, R.S. Baker, "Combining click-stream data with NLP tools to better understand MOOC completion", in 6th International Conference on Learning Analytics and Knowledge, pp. 6-14, ACM 2016.

[5]] M. Fei, D.Y. Yeung, ―Temporal models for Predicting Student Dropout in Massive Open Online Courses‖, 2015 IEEE International Conference on Data Mining Workshop (ICDMW).

[6] Fisnik Dalipi1,2, Ali Shariq Imran3 , ZenunKastrati,‖ MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges‖, 2018 IEEE Global Engineering Education Conference (EDUCON), 17-20 April, 2018, Santa Cruz de Tenerife, Canary Islands, Spain (PP1007-1014)

[7].D.F.O.Onah1 , J.Sinclair1 , R.Boyatt1, ―DROPOUT RATES OF MASSIVE OPEN ONLINE COURSES: BEHAVIOURAL PATTERNS‖, 1 The University of Warwick (UNITED KINGDOM).

[8] Heather B. Shapiro , Clara H. Lee , Noelle E. Wyman Roth , , Kun Li , Mine Çetinkaya-Rundel , Dorian A. Canelas ,‖ Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers‖ , Computers & Education, Elseiver, 2017 pp.no 35-50.

[9] L. Yan, S. Bowel, (2013) MOOCs and Open Education: Implications for Higher Education. http://publicaions.cetis.org.uk/2013/667. (Consulted 10 November 2017).

[10] Y. Belanger, and J. Thornton, (2013). Bioelectricity: A quantitative approach. Duke University'sFirst MOOC.
https://dukespace.lib.duke.edu/dspace/bitstream/handle/10161/6216/Duke_Bioelectricity_MOOC_Fall2012.pdf?sequence=1.