Education and Information Technologies (2019) 24:1973–1994
https://doi.org/10.1007/s10639-018-9829-9

Deciphering the attributes of student retention in massive open online courses using data mining techniques

Shivangi Gupta[1,2] · A. Sai Sabitha[1]

Received: 7 March 2018 / Accepted: 22 October 2018 / Published online: 31 October 2018

Abstract

Aimed at a massive outreach and open access education, Massive Open Online Courses (MOOC) has evolved incredibly engaging millions of learners' over the years. These courses provide an opportunity for learning analytics with respect to the diversity in learning activity. Inspite of its growth, high dropout rate of the learners', it is examined to be a paramount factor that may obstruct the development of the e-learning platforms. Fabricating on the existing efforts of retaining learners' engagement prior to learning, the study explores to decipher the attributes of student retention in e- learning. The study proposes a clear rationale of significant attributes using classification algorithms (Decision Tree) in order to improve course design and delivery for different MOOC providers and learners'. Using the three MOOC datasets, this research work analyses the approach and results of applying the data mining techniques to online learners', based on their in-course behaviour. Finally, it predicts the attributes that lead to minimise attrition rate and analyse the different cohort behaviour and its impacts for dropouts using data mining technique. It focuses to build a more integrated environment for these learners'.

Keywords Accuracy · Classification · Decision tree · Dropout rate · e-learning · Learning technology · Prediction · Software agents

✉ Shivangi Gupta
shivangi.gpt35@gmail.com

A. Sai Sabitha
saisabitha@gmail.com

1 Department of Information Technology, Amity School of Engineering and Technology, Amity University Campus, Sector-125, Noida, UP 201313, India

2 Delhi, India

1 Introduction

The Massive Open Online Courses (MOOCs) are a recent introduction to online education. It allows people with different interest to enrol for courses, free of cost or at substantially low cost and allowing millions of enrolments in these online courses which is the mainstream since 2012. Many educational Universities/ Institutions showed considerable effort in providing the course material to learners'. The faculties of various top universities like Harvard, MIT, and Stanford include these courses in their curriculum. Short video lectures are prepared by the teachers and these have been uploaded primarily for teaching to the learners'. Other activities such as assignments, discussion forum, online quizzes and chapter wise course structures are provided by these online courses. The professors and students are attracted because of the benefits provided by these online courses. In contrary, some online courses have shown better rates, but the completion of the courses remained at low rates. For example, some joined the course merely to observe one topic and then leave. This should not be considered as a dropout, because the MOOC success does not totally depend on the high rate of enrolees or those who have successfully completed the course. Rather, it is also built upon learners' who have successfully achieved what they want from the course (Arora et al. 2017). The most important thing is the skill and knowledge that a learner has gained at the end of the course.

Certain problems existed with the development of these online courses and the persisting issue is the high dropout rates. It has been observed that 70% of the learners' dropped out of the courses without completion and if not considered, the dropout rates can go beyond 90% (Boyer and Veeramachaneni 2015). It was foreseen that online learning might agitate the education field, but it has not appeared yet, although, MOOC has expanded significantly (Shah 2015), still is condemned for low completion rates. These online courses might be seen as free courses for learners', but for those organisations or institutions who design the course structure for online learning involve costs and is not free. The cost may vary from US$50000 – US$100000 (Bates 2013). Therefore, the high dropout rates adversely affect the institution, which develops the design for online courses.

In the past few years, online learning has become a major source of learning amongst learners', and in specific, higher education is adopting the technology of virtual online learning (Refer Fig. 1). There are a large number of e-learning providers, providing different course structure and all the providers are facing a similar problem of high attrition rates. It is observed that many people who enrol in a particular course don't visit or they never come back after two to three visits (Rodriguez 2012). Also, it has been observed that some learners' dropout of the course, even if they join the course with the intention of learning and completing (Khalil and Ebner 2014).

Based on the research, it has been observed that, not even 20% of the learners' have completed the course successfully. Researchers, regard this issue as a lack of personal motivation and commitment of individuals (Liyanagunawardena et al. 2014). At present, there are various quality factors and extensions that reside in e-learning field. Yet, these factors do not give effective results when applied to online courses. Meanwhile, there is a lack of research to predict the factors that are really affecting the quality of these online courses.

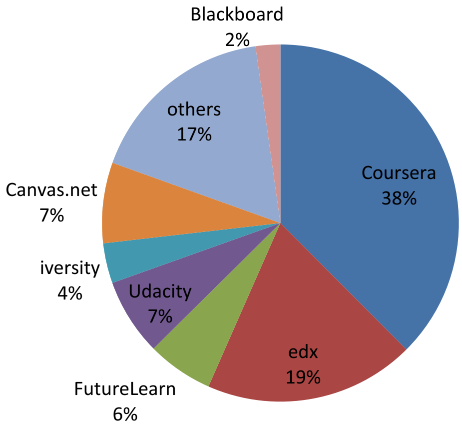


Fig. 1 MOOC providers of online learning

Inspite of a tremendous growth in MOOC design, development and delivery some attributes need improvement in providing better training for teachers, better content, design and understanding some of the important factors that affect the student dropout from the course. Understanding and improving the important attributes of these online courses can help retain the learners' in the course. Adequate software modules can be built to reduce the attrition of learners'. Therefore, significant research is required to understand the nature of learners' to improve the quality of e-learning and deliver good content. Adaptive learning technologies are needed to be implemented to analyse the engagement of users with online learning environment thereby making it enormously successful.

This research work attempts to understand the important attributes of online learning environment. It aims to understand the student's intention in learning environment and learners' involvement throughout the course in order to build an immensely strong MOOC environment. An algorithm is a set of instructions that are to be followed especially by computer, in problem – solving and calculations (Yousef et al. 2014). Therefore, the most appropriate algorithm is recognised by finding the accuracy of the different classification algorithm. The Decision Tree came was considered the most appropriate algorithm, as it showed the highest accuracy. Decision tree is a map (tree-like structure) that depicts the possible outcome of making a decision (Bharara et al. 2018). This research work focuses on Student perception features and other attributes to understand the learners' dropout from MOOC, so that one-course with similar characteristics can be adopted by other courses with similar characteristics.

2 Literature review

In order to find the important factors affecting the online learning environment plenty of research works has been carried out. Many quality factors have been discovered to have an efficacious e-learning environment. It has not been proven totally effective because of the unique features of different e-learning platforms (Hegyesi et al. 2017). Various machine learning models such as Support Vector Machine, Logistic Regression were tested. They showed 80% accuracy and predicted dropout, based on the training dataset (Sharkey and Sanders 2014). During the past few years, researchers have made massive progress in identifying the cognitive processes important in learning. Cognitive processes worked upon revealing the emotional state of learners' and the interconnection between emotion and efficient learning (Shen et al. 2009).

2.1 Data mining approach in MOOC

Data mining is a field of intersection of computer science and statistics used to discover patterns and extract the useful information from the dossier of data and mould it into an understandable structure for future use (Agarwal 2013). Online courses were developed as a substitute to educational platform that allow learners' from different locations to access the same quality of material through internet. Educational Data Mining (EDM) has also contributed to support decision makers by withdrawing knowledge of learning processes. It has been used to strengthen the e-learning design by enhancing teaching strategies. Learning Analytics (LA) also contributed in enhancing the quality of education (Al-Shabandar et al. 2017). EDM and LA were designed with the purpose to understand learners' involvement with MOOC efficiently. LA helped in recognising the dropout students. Studies have been conducted on how motivation impacts in learners', course learning, retention and completion (Sooryanarayan and Gupta 2015). Table 1 show the research work conducted in online learning. It was observed that some previous work proposed to use quality metrics as quality aspect to be considered in the MOOC. It was strongly opposed stating that it cannot be used because many e-learning environments has pedagogical features which can't be validated by online learners' (Yousef et al. 2014).

Some of the reasons for their attrition could be as follows:

Table 1 Research work in massive open online courses

S. no.	Authors	Year	Application	Techniques used
1.	Huang et al.	(2017)	Explored interaction between Learners' and Video Content	VideoMark Analytical Approach for Video Based Learning
2.	Sunar et al.	(2016)	How Learners' Sustain Engagement through Interaction in MOOC	MOOC Case Study: Social Network Analysis Technique
3.	Arora et al.	(2017)	Learner Group in Massive Open Online Courses	K-means Clustering to obtain clusters of learners' having analogous interaction
4.	Mulik et al.	(2016)	Identifying Key Determinants for MOOC Acceptance	Acquainted Technology Acceptance Model (TAM) to find determinants for MOOC acceptance by learners'
5.	Sabitha et al. 2016	(2016)	Converging learning objects with open educational resources	Naive Bayes Approach
6.	Sooryanarayan and Gupta	(2015)	Learner Motivation on MOOC Preferences	Logit Regression

- Shortage of time
- Inadequacy of motivation
- Absence of interaction
- Unable to relate the skills and knowledge being taught in MOOC

2.2 Use of software agents in MOOC

Educational Engineers see the web as a big information system that provides learning resources to the world population for good learning experience (Sunar et al. 2016). MOOC prominence was shown in 2012 and 2013, when different organisations came forward to provide MOOC course design. Some provided free services while others provided paid services for certification or a small fee for selective courses. Software agents are similar to computer systems to which a task can be assigned (Bassi et al. 2014). In online learning, software agents can analyse the material on MOOC platform, behaviour of the participant and the system to intelligently upgrade the delivery and assessment system of e-learning platform. Introducing the agents into an interactive multimedia system can help in increasing the effectiveness of multimedia background (Kaveri et al. 2016). The different agents used so far are Pedagogical agents (like characters used for interaction in e-learning environment), Web agents (software system designed to perform searching and filtering functions), Learner's agent (they take performance decisions to create efficient learning environment to learners') and Mixed agents (provides support for completion of task in an application). Initially, these were used in Intelligent Tutoring Systems (ITS). It helps the learners' in completion of the task without human intervention and later on in the Virtual Learning Environment (web – based platform for digital aspect of course study) (Machado and Ruiz 2017).

Comprehensive use of the internet has resulted in the evolution of new pedagogies and learning patterns. In 2005, a new theory was proposed by (Siemens 2005) known as "connectivist" learning theory. First these online courses were based on the Connectivism in which the learning was done through making meaningful connections between knowledge, information systems and ideas. This was done to trigger a conversation during the course. It was found by Rodriguez (2012), that e-learning platform adopted cognitive and social constructive approach. E-learning is the adoption of information and communication technology that enables people to read anytime and anywhere (Sandanayake and Madurapperuma 2013). There are different methods in which a trainer can communicate with online learners' (Zhou et al. 2016).

2.3 Review of work done in MOOC

Several researches have been carried out in pursuance to find out the different factors affecting online learning. Walker and Loch carried out an empirical research, in which they carried out a survey for people who participated in online learning through different platforms such as twitter, e-mails or any other personal networks (Gamage et al. 2015). They concluded that learners' bought out the need for technological aspects such as video, assignments and discussions. The enhancement in the above mentioned parameter can help retain learners' in e-learning platform.

Adamopoulos (2013) confederated the Grounded Theory (GT) in MOOC, which carried a unique analysis using user feedback to determine the factors that built enormous online learning platform. The research of GT can also be used in quantitative studies. Later, Strauss came out with the research of human involvement which is more important than the passive enroler. GT was based on the massive participation of learners' and depicted behaviours and patterns of the learners' (Adamopoulos 2013). This work also gave a vast range of reasons for immense dropout.

Schaffer et al. (2016) performed the visual network analysis to understand the behavioural features that best predicted the student attrition with different courses. They focused on visualisation of student data and concluded that those students who never received the response on discussion forum were likely to attrite. Analysing the dropout from these online courses, Schaffer provided guidance to MOOC trainer and engineer for better designing of these courses. Yang identified that intervening to guide learners' those who are facing difficulty, can reduce attrition rates to a great extent (Rosé et al. 2014). Populous numbers of learners' with unanswered posts were identified with the help of visualisation.

Researchers have analysed how the peers can help with retention of learners' in online courses. They studied, how peers form bonds and can help students remain active throughout the course. During adapting the Technology Advancement Model, it was established that a particular assumption as perceived usefulness and perceived ease of use to determine acceptance behaviour of computer system (Mulik et al. 2016). One of the attrition factors was that whether learners' are able to correlate the technology with their skill and understanding or not. This work analysed the behaviour intentions to examine the acceptance of technology by MOOC learners'.

Online courses had created various opportunities in the field of education and most importantly for organisational stakeholders. For enhancement of these online courses, additional features of video lectures were presented by Huang et al. (2017). VisMOOC (Shi et al. 2015) was proposed to analyse learners' online learning behaviour based on the video clickstream data (Gallén and Caro 2017). Table 2 shows different data mining techniques used to analyse MOOC attrition.

Online courses have been proven beneficial to almost every person irrespective of their need, but the high dropout rates are hindering MOOC development (Kloft et al. 2014). The Statistical models have been developed to overcome this problem. Chen et al. (2016) noticed that these statistical models could predict dropout, but the people may not understand the reasons behind the predicted result. Moreover, it would be difficult for MOOC technologists to bring the modifications in online courses for better delivery and assessment. So, he developed a Dropoutseer, a visual analytics system, which helps educational trainers and developers to understand the critical features of dropout (Wu et al. 2016). This helped them to accomplish good models with effective performance.

3 Methodology

The research aims to determine important elements that influence the online learners' perspective to dropout. The study is based on one of the classification algorithms, i.e., Decision Tree to analyse the social reality and contextual

Table 2 Research work done with different technique to improve MOOC Attrition

S. no.	Authors	Year	Application	Techniques used
1.	Bassi et al.	(2014)	Analyse challenges in MOOC Design, Delivery and Assessment	Usage of Software Agents to overcome identified reasons in MOOC dropout
2.	Al-Shabandar et al.	(2017)	Behavioural Patterns were compared to Predict Learners' Retention in Course	To improve the accuracy of Classifier Models, Machine Learning Algorithms were applied
3.	Bharara et al.	(2017)	Application of learning analytics	Clustering Data Mining for Students' Disposition Analysis
4.	Gallén and Caro	(2017)	Data accumulation of MOOC Knowledge project	Knowledge Discovery in Database, clustering
5.	Schaffer et al.	(2016)	Predicting Attrition by examining the effect of different Course Structure on Learners'	Distinct Informational Visualisation of Student Network Data
6.	Chen et al.	(2016)	Visual Analytics System to help Educational Instructors to understand the reason for Dropout	DropoutSeer: Visual Analytics Predictive Modeling
7.	Wu et al.	(2016)	Analyse large group of people at each step	EgoSlider: To visualise system that explores, compare and analyse social network
8.	Boyer and Veeramacha-neni	(2015)	Predicting learners' those who are likely to stop engagement in MOOC	Data recorded of Learners' to build prediction Model
9.	Gamage et al.	(2015)	To identify Student's Behavior and requirements in MOOC	Grounded Theory, methodology to analyse Student's behaviour and requirements in MOOC
10.	Khalil and Ebner	(2014)	MOOCs attainment rates and possible methods to improve retention-A literature review	Focuses on reasons that are responsible for Student's withdrawals
11.	Rosé et al.	(2014)	Explore Dropout Behaviour of Learner in Massive Open Online Courses	Survival Model to account the impact of Social Factors on Attrition
12.	Kloft et al.	(2014)	Predicting MOOC using Machine Learning Methods	Classification based on Clickstream data, Machine Learning Algorithm

importance in online learning environment. Data collection and extensive pre-processing was performed on the three datasets of MOOC in order to find a technique that is best suited for analysing dropout behaviour in online courses. The important factors for dropout were predicted by generating a decision tree of three different MOOC datasets using Programming.

The different MOOC providers may have different attributes in their respective course design. The other attributes responsible for learners' dropout from online courses has been found out by carrying a study on different MOOC datasets. This has been done so that one course cohort can be used to accommodate the design of the course for another cohort. Learning is the significant facet of several education systems.

This research work follows the following steps:

Step 1: Data Collection.
Step 2: Analysing attributes and features of the dataset.
Step 3: Extensive Preprocessing and data cleaning.
Step 4: Determining the best classification algorithm by finding out accuracy on each dataset.
Step 5: Normalisation of the Dataset.

Therefore, in Section 4 Decision Tree is used to identify the predominant factors so that different MOOCs can be mapped to standard learning approach by focusing on identified predominant factors.

Step 6: Experimental setup.
Step 7: Decision tree formation and analysis.

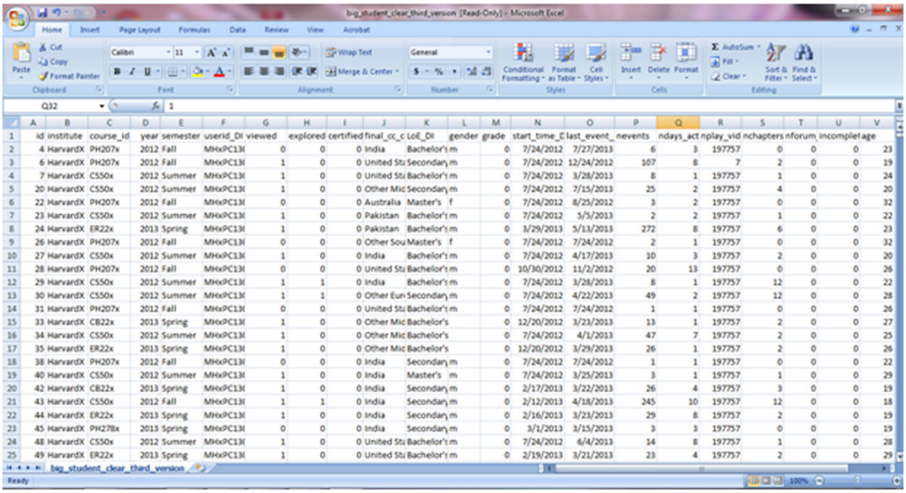
3.1 (Step 1): Data collection

This research work was carried based on the three datasets of MOOC. The datasets were collected from Kaggle.com and Dataverse.

The first dataset used is "big_student_clear_third_version" (Refer Fig. 2) from Kaggle.com by alishanmustafa (Kaggle 2017a). This dataset contains 4, 16, 992 rows and 21 columns and the information of students of MIT and Harvard enroled in MOOC in three semesters (Fall, Spring and Summer).

The second dataset used was another MOOC dataset named as "cs_mitx" (Refer Fig. 3) from Kaggle.com by Dan Ofer (Kaggle 2017b). This dataset comprises of 59, 280 rows and 23 columns containing information about the learners' of MIT enroled in MOOC courses.

The third data set used in this paper was the MOOC dataset named as "Harvard-MIT Person-Course De-identified Dataset", (Refer Fig. 4) Version 2.0 from Harvard



id	institute	course_id	year	semester	userid_DI	viewed	explored	certified	final_cc_c	LoE_DI	gender	grade	start_time_D	last_event_	nevents	ndays_act	nplay_vid	nchapters	nforum	incomplet	age
4	HarvardX	PH207x	2012	Fall	MHxPC13	0	0	0	India	Bachelor's	m	0	7/24/2012	7/27/2013	6	3	197757	0	0	0	23
6	HarvardX	PH207x	2012	Fall	MHxPC13	1	0	0	United Sta	Secondar	m	0	7/24/2012	12/24/2012	107	8	7	2	0	0	19
7	HarvardX	CS50x	2012	Summer	MHxPC13	1	0	0	United Sta	Bachelor's	m	0	7/24/2012	3/28/2013	8	1	197757	1	0	0	24
20	HarvardX	CS50x	2012	Summer	MHxPC13	1	0	0	Other Mic	Secondar	m	0	7/24/2012	7/15/2013	25	2	197757	4	0	0	20
22	HarvardX	PH207x	2012	Fall	MHxPC13	0	0	0	Australia	Master's	f	0	7/24/2012	8/25/2012	3	2	197757	0	0	0	32
23	HarvardX	CS50x	2012	Summer	MHxPC13	1	0	0	Pakistan	Bachelor's	m	0	7/24/2012	5/5/2013	2	2	197757	1	0	0	22
24	HarvardX	ER22x	2013	Spring	MHxPC13	1	0	0	Pakistan	Bachelor's	m	0	3/29/2013	5/13/2013	272	8	197757	6	0	0	23
26	HarvardX	PH207x	2012	Fall	MHxPC13	0	0	0	Other Sou	Master's	f	0	7/24/2012	7/24/2012	2	1	197757	0	0	0	32
27	HarvardX	CS50x	2012	Summer	MHxPC13	1	0	0	India	Bachelor's	m	0	7/24/2012	4/17/2013	10	3	197757	2	0	0	20
28	HarvardX	PH207x	2012	Fall	MHxPC13	0	0	0	United Sta	Bachelor's	m	0	10/30/2012	11/2/2012	20	13	197757	0	0	0	26
29	HarvardX	CS50x	2012	Summer	MHxPC13	1	1	0	India	Bachelor's	m	0	7/24/2012	3/28/2013	8	1	197757	12	0	0	22
30	HarvardX	CS50x	2012	Summer	MHxPC13	1	1	0	Other Eur	Secondar	m	0	7/24/2012	4/22/2013	49	2	197757	12	0	0	28
31	HarvardX	PH207x	2012	Fall	MHxPC13	0	0	0	United Sta	Bachelor's	m	0	7/24/2012	7/24/2012	1	1	197757	0	0	0	26
33	HarvardX	CB22x	2013	Spring	MHxPC13	1	0	0	Other Mic	Bachelor's		0	12/20/2012	3/23/2013	13	1	197757	2	0	0	27
34	HarvardX	CS50x	2012	Summer	MHxPC13	1	0	0	Other Mic	Bachelor's		0	7/24/2012	4/1/2013	47	7	197757	2	0	0	25
35	HarvardX	ER22x	2013	Spring	MHxPC13	1	0	0	Other Mic	Bachelor's		0	12/20/2012	3/29/2013	26	1	197757	2	0	0	26
38	HarvardX	PH207x	2012	Fall	MHxPC13	0	0	0	India	Secondar	m	0	7/24/2012	7/24/2012	1	1	197757	0	0	0	22
40	HarvardX	CS50x	2012	Summer	MHxPC13	1	0	0	India	Master's	m	0	7/24/2012	3/25/2013	3	1	197757	1	0	0	29
42	HarvardX	CB22x	2013	Spring	MHxPC13	1	0	0	India	Secondar	m	0	2/17/2013	3/22/2013	26	4	197757	3	0	0	19
43	HarvardX	CS50x	2012	Fall	MHxPC13	1	1	0	India	Secondar	m	0	2/12/2013	4/18/2013	245	10	197757	12	0	0	18
44	HarvardX	ER22x	2013	Spring	MHxPC13	1	0	0	India	Secondar	m	0	2/16/2013	3/23/2013	29	8	197757	2	0	0	19
45	HarvardX	PH278x	2013	Spring	MHxPC13	0	0	0	India	Secondar	m	0	3/1/2013	3/15/2013	3	3	197757	0	0	0	19
48	HarvardX	CS50x	2012	Summer	MHxPC13	1	0	0	United Sta	Bachelor's	m	0	7/24/2012	6/4/2013	14	8	197757	1	0	0	28
49	HarvardX	ER22x	2013	Spring	MHxPC13	1	0	0	United Sta	Bachelor's	m	0	2/19/2013	3/21/2013	23	4	197757	2	0	0	29

Fig. 2 Screenshot of dataset "big_student_clear_third_version"

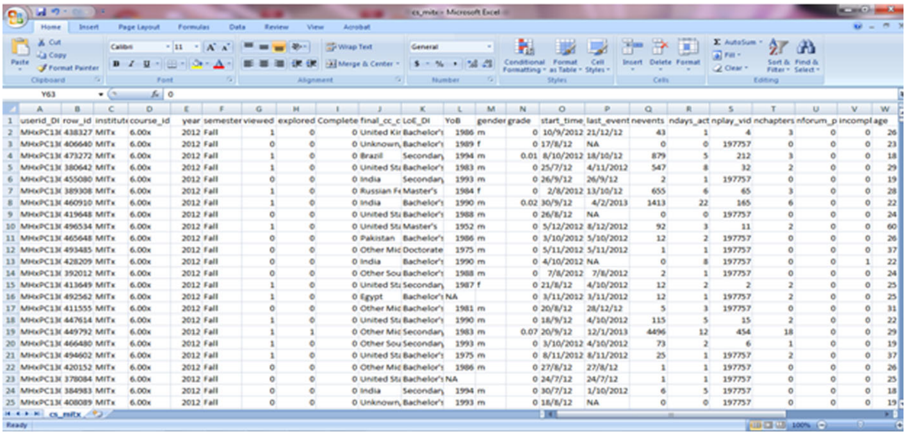


Fig. 3 Screenshot of dataset "cs_mitx"

Dataverse (Dataverse 2014). The data set incorporates 6, 41, 139 rows and 20 columns. The data set contains information of the total number of people of MIT or Harvard enroled in online courses (edx platform) and the attributes show activities performed by them during the course.

3.2 (Step 2): Analysing attributes and features of the dataset

Online courses are provided by different service providers, educational institutions and technologist. Therefore, designing the structures and attributes might be different for each of them. Similarly, the above three MOOC datasets (Refer Figs. 2, 3 and 4) considered have different attributes. The experimental analysis would require important input parameters to predict important dropout attributes. So, this study involves analysis of each attribute and selected the relevant attribute for experimental analysis.

Table 3 shows the dataset attributes and their description to understand the role of each attribute in the MOOC platform. This work analysed that, learners' recognition

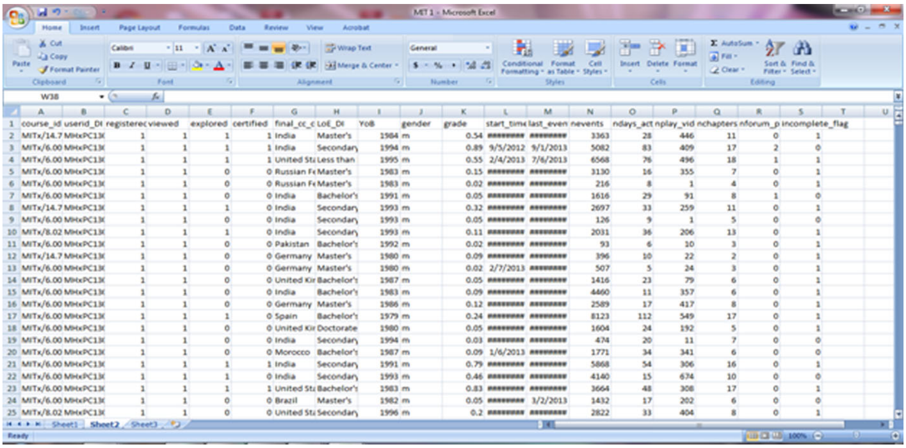


Fig. 4 Screenshot of dataset "Harvard-MIT Person-Course De-identified Dataset"

Table 3 List of attributes of dataset "big_student_clear_third_version", "cs_mitx", "Harvard-MIT Person-Course De-identified Dataset"

Attributes	Description
institute	Name of the institute that the learner belongs to
course_id	Id associated with the course
year	Year of Enrolment
semester	The Semester in which the learner is enroled
userid_DI	Id associated with the user
viewed	Number of times the learner viewed the course
explored	Number of times the learner explored the course
certified	The Learner is certified or not at the end of the course
final_cc_cname_DI	Name of the country the learner belongs to
LoE_DI	Level of Enrolment (Bachelor's or Master's)
gender	Gender of the learner
grade	Grades obtained in the course
start_time_DI	Start Date of the course
last_event_DI	End Date of the course
nevents	Learner participation in the total number of events
ndays_act	Total days the learner was active during the course
nplay_video	Total videos played by the learner
nchapters	Total chapters explored by the learner
nforum_posts	Total number of posts generated by the learner
incomplete_flag	Events that were marked for later, but not completed
age	Age of the learner
YOB	Learner's year of birth

features are the most important to understand the learner's dropout from the course (Refer Section 4). Recognition features such as Certified, nEvents (number of events a learner was involved), Days Active, Played Video, nChapters (number of chapters explored).

3.3 (Step 3): Extensive preprocessing and data cleaning

The datasets included some missing values which might affect the results hence, all such values including null values were removed from the datasets. Normalisation (pre-processing stage that helps in finding a new range from the existing range) of data was done by setting new max to 10 and new min value to 1. The Datasets were divided into distinguishable parts to make it easier to work with. The first dataset used "big_student_clear_third_version" was divided into MIT and Harvard, Males and Females and further, the data was categorised on the basis of semester (Fall, Spring, Summer) (Refer Fig. 5). A total of 24 excel sheets were prepared to find the accuracy of classification technique to predict the important attributes that play of a vital role in the learners' dropout from online courses.

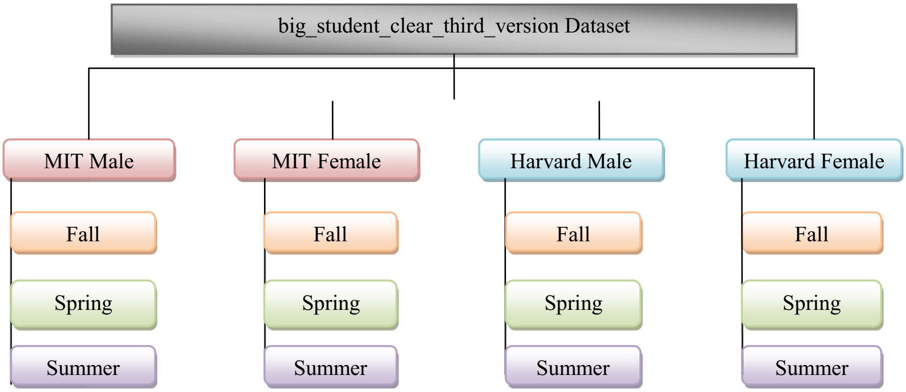


Fig. 5 Divisions of the dataset into 12 excel sheets to determine the best classification algorithm

The second dataset named "cs_mitx" was also transformed by eliminating all the null values. This dataset contained all learners' who enroled in the online course during the Fall semester containing data of all learners' from MIT. Spring and Summer enrolment was not present hence, the dataset was characterised as Male and Female (Refer Fig. 6).

The Third "Harvard-MIT Person-Course De-identified Dataset" contains information of the total number of people of MIT or Harvard enroled in the online courses (edx platform) and similarly this data was cleaned by removing the missing values from the dataset. The dataset was categorised into three sets as Student Behaviour (registered, viewed, explored, certified, gender), Student Perception (number of events, certified or not, days active, videos played, number of chapters explored) and Student's Assertion (course id, user id, year of birth, gender, grade and nforum post). According to the previous work done in the MOOC, Student perception attributes were taken into consideration as the predicting factors for learners' dropout from the online courses. The dataset was divided into MIT and Harvard and further subdivided into Male and Female (Refer Fig. 7).

3.4 (Step 4): Determining the best classification algorithm by finding out accuracy on each dataset

The Decision Tree, Random Forest, K nearest neighbor and naïve Bayes techniques were applied on each distinguished dataset (Refer Section 3.3) using Rapid Miner tool. Validation was applied to separate the dataset into training and testing (Bharara et al. 2017). A total of 24 excel sheets were prepared from three datasets (Refer Section 3.1), each sheet was tested across the classification algorithms (Random Forest, Decision

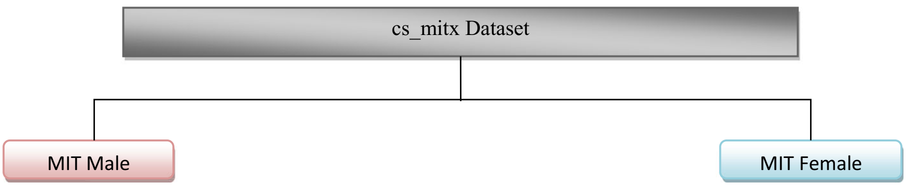


Fig. 6 Divisions of the dataset into 2 excel sheets to determine the best classification algorithm

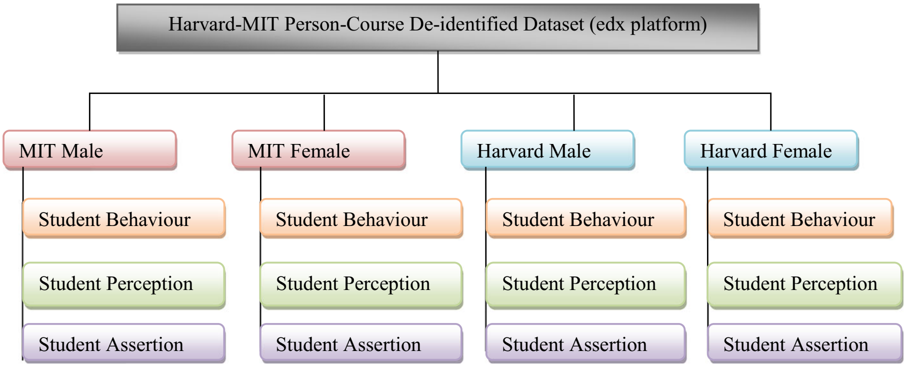


Fig. 7 Divisions of the dataset into 12 excel sheets to determine the best classification algorithm

Tree, KNN, and Naïve Bayes) and accuracy was noted for each sheet. The Decision Tree showed relatively highest accuracy amongst all algorithms (Refer Table 4). Therefore, the Decision Tree was chosen to predict the dropout parameters in the online courses.

The Decision Tree algorithm forms tree using concrete values. It uses the concept of divide and conquer to form the Decision Tree (Onah et al. 2014). Make sure to make the values concrete, if the values are not concrete it may not generate accurate results.

In the Random Forest classification algorithm the cluster of the Decision Tree is seen. The tree votes are given to each tree to resemble classification (Sabitha et al. 2016). All trees are developed at its maximum length possible. A tree with more number of votes are chosen as forest.

The K Nearest Neighbor (KNN) works practically and is a smooth working classification algorithm. It uses the concept of distant function for mapping of sample and classes (Castro and Tsuzuki 2015). KNN is the best suitable algorithm in case of application objects having multiple labels. It gives good accuracy and reliable performance.

The Naïve Bayes classification algorithm works on the concept of Bayes' theorem and assumes that all features are independent. It works entirely for the real time problems. It works on the discrete and continuous values. In this the performance would be degraded if one attribute is dependent on another attribute.

3.5 (Step 5): Normalisation of the dataset

One of the important preprocessing techniques of the dataset is normalisation. Before normalisation there is a need to remove irrelevant attributes and missing values. Selection of the feature is the most important step of preprocessing. The main focus is to select a subclass of feature that can be used as an input data and decrease the unsuitable data. This step helps in increasing the accuracy of implementation done on the given dataset. The Focus is on Student Perception features (as shown in Table 5) for early prediction of dropout reason from the online courses. In this the learners' performance is converted from numerical values to nominal values. The Dataset is divided into three intervals (High, Medium and Low) based on the total number of active days like values between 0 and 50 that fall under low interval, values between 51

Table 4 Determining accuracy of different classification algorithms

Dataset	Institutes	Semester	Gender	Decision tree	Random forest	KNN	Naïve bayes
big_student_clear_third_version Dataset	Harvard	FALL	Male	98.99	96.95	92.29	98.26
			Female	98.99	97.07	96.32	98.37
		SPRING	Male	98.99	96	94.01	96.51
			Female	98.91	97.72	85.99	88.43
		SUMMER	Male	98.88	98.07	97.27	99.97
			Female	98.99	98.93	94.32	98.87
	MIT	FALL	Male	99.97	98.39	91.36	96.78
			Female	99.94	98.39	95.24	96.4
		SPRING	Male	99.97	97.9	91.66	95.97
			Female	98.26	97.28	87.43	95.79
		SUMMER	Male	98.99	98.82	97.91	97.06
			Female	99.96	98.75	88.15	97.95
cs_mitx Dataset	MIT	FALL	Male	97.61	74.11	67.96	85.5
			Female	82.89	83.19	64.41	88.71
Harvard-MIT Person-Course De-identified Dataset (edx platform)	Harvard	STUDENT BEHAVIOR	Male	64.62	60.91	32.32	62.96
			Female	95.09	91.34	89.41	90.1
		STUDENT PERCEPTION	Male	64.74	64.4	62.81	53.14
			Female	93.62	92.52	91.46	90.03
		STUDENT ASSERTION	Male	67.28	64.34	64.19	66.89
			Female	99.95	94.71	88.91	82.16
	MIT	STUDENT BEHAVIOR	Male	82.89	82.17	78.49	82.59
			Female	82.59	82.17	78.49	82.59
		STUDENT PERCEPTION	Male	88.98	88.04	85.39	85.59
			Female	87.41	87.31	78.56	84.06
		STUDENT ASSERTION	Male	98.04	95.91	96.68	94.44
			Female	98.74	98.39	95.77	96.28

to 100 fall under medium interval and values between 101 and 150 fall under high interval. The other features are also normalised, such as nEvents, days active, video played and nChapters using the formula $z_i = ((v\text{-}x_{min}) / (x_{max} - x_{min}) * (y_{max} - y_{min}) + (y_{min}))$. Table 5 shows the normalised value of the selected attributes.

4 Experimental analysis

4.1 (Step 6): Experimental setup

The best suitable data mining technique for predicting important factors in learners' dropout is the Decision Tree. The Decision Tree works on divide and conquer

Table 5 Normalisation of dataset

CE	NE	DA	PA	NC	AE
0	1.069097	1.672897	1.021997	1.6	3.442857
1	1.0161	1.168224	1.003142	1.6	4.728571
0	1.18381	1.336449	1.336243	1.6	3.957143
1	1.584973	2.009346	1.521648	2.2	5.757143
1	1.003354	1.084112	1.01257	1	3.828571
1	1.023479	1.252336	1.01257	2.2	4.857143
0	1.83788	3.186916	1.782472	9.4	4.214286
1	1.015429	1.252336	1.01257	1.6	3.957143
1	5.578488	4.785047	2.759777	9.4	4.471429
0	6.135957	7.056075	4.946927	9.4	4.085714
0	1.016771	1.336449	1.01257	1	5.757143
1	1.077147	1.420561	1.062849	1.6	3.7
0	1.046288	1.168224	1.009427	1.6	4.857143
1	1.042934	1.252336	1.02514	1	3.957143
1	2.69119	3.186916	1.430517	4	4.085714
0	1.366279	2.093458	1.254539	2.8	4.6

technique. It gives the good performance for concrete values. The Decision Tree also showed relatively maximum accuracy on datasets, as shown in Table 4. The Decision Tree was applied on three datasets (Refer Section 3.1). The attributes of student perception were considered as defined in the Normalised dataset (Refer Table 5). The Attributes were as follows:

- CE: Certified or not
- NE: Total number of events participated
- DA: Total number of days active
- PV: Total number of videos played
- NC: Total number of chapters explored
- AE: Age of the learners’

Using R programming, in this study, the Decision Tree is generated as shown in Fig. 8. In R, the Decision Tree is used to illustrate the choices and the results are shown as Tree. The nodes signify the choices made and the edges represent the conditions as shown in Fig. 9.

4.2 (Step 7): Decision tree formation and analysis

The three Decision Trees are generated and analysed using R Studio on the basis of different parameters selected such as explored course, certified in the course or not, number of days active in the course, the number of events a learner participated, a number of videos played by learners’ and age.
> view (cs_mitxv)

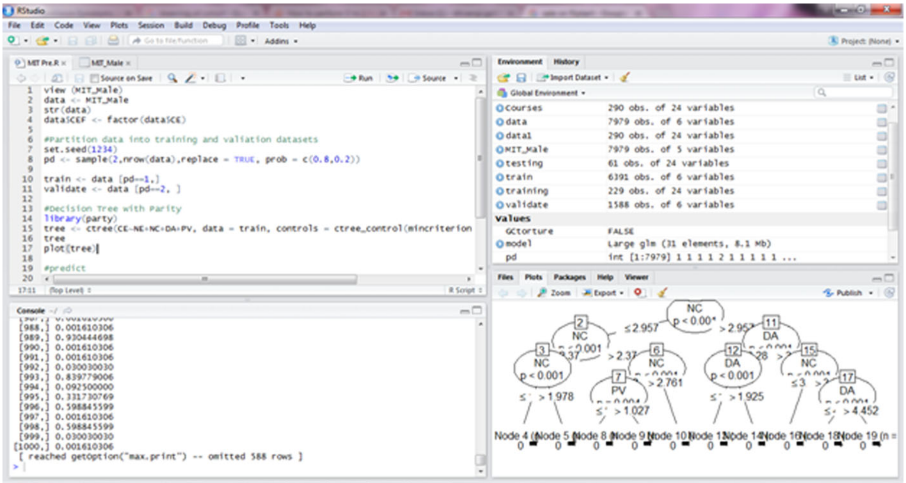


Fig. 8 Decision tree implementation in R studio

> data <– cs_mitxv
> str(data)
>data$CEF < − factor(data$CE)
#Partition data into training and validation datasets
> set.seed(1234)
> pd. < − sample(2,nrow(data),replace = TRUE, prob. = c(0.8,0.2))
> train <– data [pd==1,]
> validate <– data [pd==2,]
#Decision Tree with Parity
> library(party)

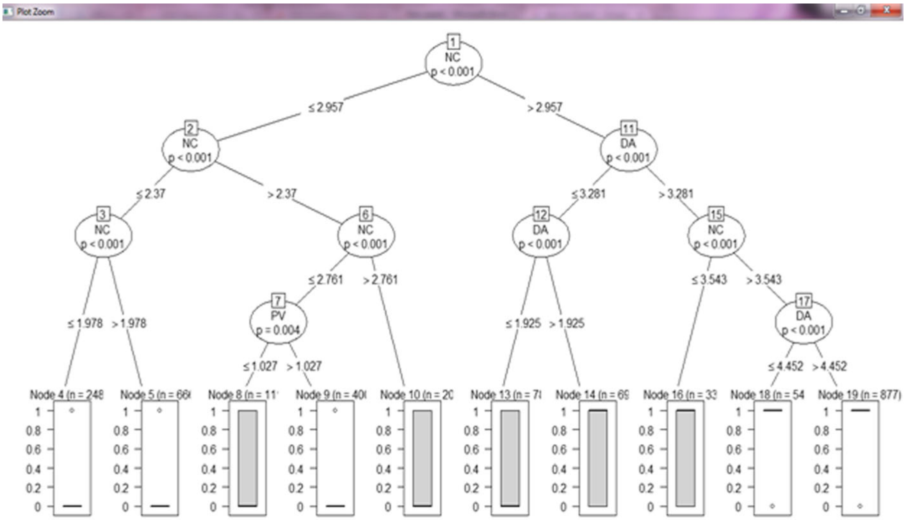


Fig. 9 Decision tree using attributes of Harvard-MIT Person-Course De-identified Dataset (edx platform)

> tree <− ctree(CE~NE + AE + NC + DA + NV, data = train, controls = ctree_control(mincriterion = 0.99, minsplit = 500))
> tree
> plot(tree)
#predict
predict(tree,validate)

In these Decision Trees (Refer Figs. 9, 10, and 11), it is interpreted that the nChapters (number of chapters explored), Days Active (total days a student was active), nEvents (number of events a student participated in), played video (number of videos played) are the most important attributes for predicting learners' dropout from the MOOC. The range values between zero to two (0–2) are considered low level interval, two to three (2–3) are considered as medium level interval and above 3 the range values are considered under high level interval. All the three datasets used, shows similar features and analysis of one course cohort that can be used to adopt the design of other similar course for another cohort. The Successive similar cohorts are shown in this analysis.

Table 6 shows the parameters that came up in all the three datasets. The tick mark and cross mark indicates the importance of the attribute to predict the attrition rates, in that particular dataset. It shows that viewed (Course Viewed by the learner), Year of Birth (Birth year of learner) and Age (Age of the learner) are not the important attributes to predict Learners' dropout from the MOOC. These attributes cannot help in improving the course structure and delivery assessment. However, NE (Total number of events, a learner participated during the online course), Days Active, Played Video and the Number of Chapters explored are shown as important attributes in prediction of the dropout from the course. The importance of each parameter is shown in Fig. 12. The NC (number of chapters) and the DA (Days Active) are some of the important attributes to be considered while minimizing the attrition rates in online courses.

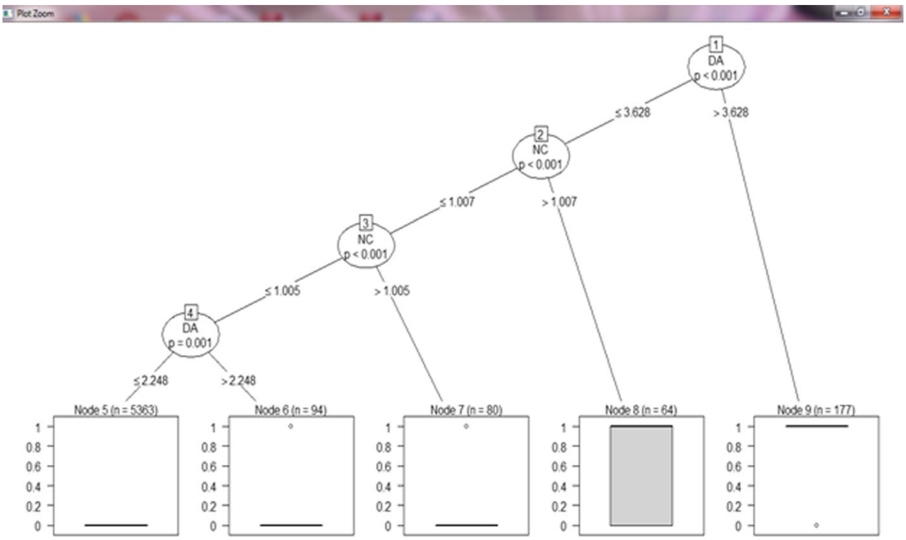


Fig. 10 Decision tree using attributes of Cs_mitx

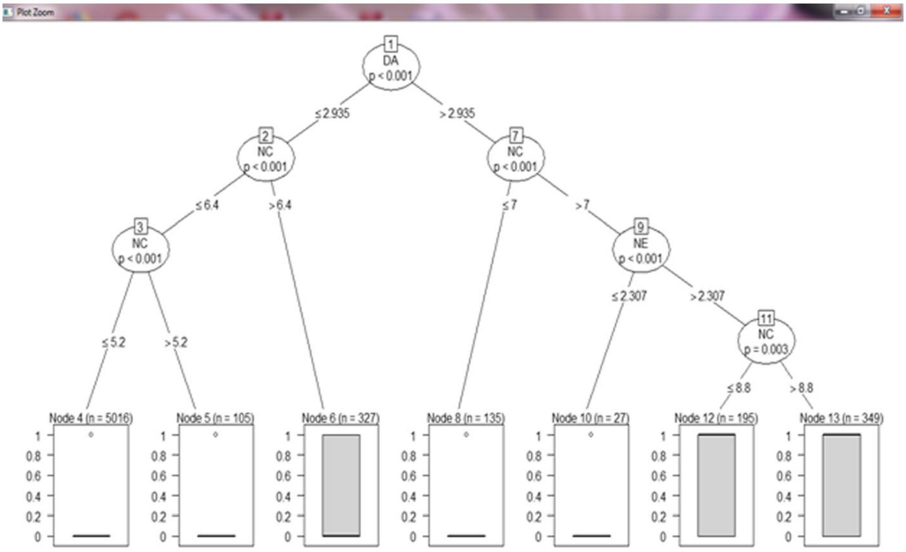


Fig. 11 Decision tree using attributes of big_student_clear_third_version Dataset

However, the PV (Played Video) and the NE (Number of Events) attributes could also be considered while improving the design structure of the MOOC to prevent the dropout of learners' and increase their engagement till the completion of course.

5 Conclusion

In judging the success of the MOOC, it is crucial to understand the learners' inclination towards the online courses and the elements impacting them. This study makes an additional benefaction to understand deeper aspects of learning to take the present understanding of the MOOC forward. This work provides a substitute way to behold the success of the online courses. The objective of this study was to perceive the influential attribute that tells the probability of learners' dropout. This work inspected the reasonable technique, i.e., the Decision Tree for finding the attributes that should be considered by educational technologists for the MOOC design and delivery. Comprehensible preprocessing was done and classification technique (Decision Tree, Random Forest, K Nearest Neighbour and Naïve Bayes) were applied and the Decision Tree algorithm was selected for analysis because it showed relatively highest accuracy amongst all the data mining techniques.

It is paramount to recognise the learners' behaviour at the initial stage to prevent them from attrition. Learners' dropout from the online courses is a dominant concern as it scales down the progress of the MOOC provider industries. This research work uses the three datasets of the MOOC ("big_student_clear_third_version Dataset", "cs_mitx Dataset", "Harvard-MIT Person-Course De-identified Dataset (edx platform)") and the Decision Tree was generated for each dataset after extensive preprocessing. The results of the analysis indicated that, the viewed (Course viewed by the learner), Year of Birth

Table 6 Important parameters to retain learners' in the MOOC

DATASET	INSTITUTE	SEMESTER	GENDER	viewed	(NE) Total Events	Days Active (DA)	Year of Birth	Played Video (PV)	Total Chapters (NC)	Age (AE)
big_student_clear_third_version Dataset	Harvard	FALL	Male	x	✓	✓	x	x	✓	x
			Female	x	✓	✓	x	x	✓	x
		SPRING	Male	x	✓	✓	x	x	✓	x
			Female	x	✓	✓	x	x	✓	x
		SUMMER	Male	x	✓	✓	x	x	✓	x
			Female	x	✓	✓	x	x	✓	x
	MIT	FALL	Male	x	✓	✓	x	x	✓	x
			Female	x	✓	✓	x	x	✓	x
		SPRING	Male	x	✓	✓	x	x	✓	x
			Female	x	✓	✓	x	x	✓	x
		SUMMER	Male	x	✓	✓	x	x	✓	x
			Female	x	✓	✓	x	x	✓	x
cs_mitx Dataset	MIT	FALL	Male	x	x	✓	x	x	✓	x
			Female	x	x	✓	x	x	✓	x
Harvard-MIT Person-Course De-identified Dataset (edx platform)	Harvard	STUDENT BEHAVIOR	Male	x	x	✓	x	✓	✓	x
			Female	x	x	✓	x	✓	✓	x
		STUDENT PERCEPTION	Male	x	x	✓	x	✓	✓	x
			Female	x	x	✓	x	✓	✓	x
		STUDENT ASSERTION	Male	x	x	✓	x	✓	✓	x
			Female	x	x	✓	x	✓	✓	x
	MIT	STUDENT BEHAVIOR	Male	x	x	✓	x	✓	✓	x
			Female	x	x	✓	x	✓	✓	x

Table 6 (continued)

DATASET	INSTITUTE	SEMESTER	GENDER	viewed	(NE) Total Events	Days Active (DA)	Year of Birth	Played Video (PV)	Total Chapters (NC)	Age (AE)
		STUDENT PERCEPTION	Male	x	x	✓	x	✓	✓	x
			Female	x	x	✓	x	✓	✓	x
		STUDENT ASSERTION	Male	x	x	✓	x	✓	✓	x
			Female	x	x	✓	x	✓	✓	x

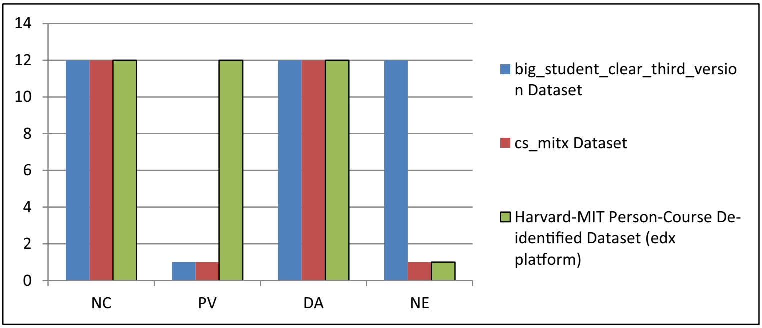


Fig. 12 Decision tree using attributes of big_student_clear_third_version dataset

(Birth Year of the learner) and Age (Age of the learner) are not significant attributes to predict the Learners' dropout from these online courses. These attributes are not helpful in determining the learning approach taxonomy. These three attributes would not be sufficient to analyse the diversified cohorts behaviour and its impact on dropout in online courses. Therefore, NE (Total Number of events, a learner participated during the MOOC course), Days Active, Played Video and the Number of Chapters explored are exhibited as influential attributes to predict the early dropout from the course.

This analysis helps to understand that one course cohort can be used to adopt the design of the course for another cohort, as this study fulfills the requirement by carrying out the analysis on three different MOOC datasets. It can be used to identify the fuzziness amongst the enrolers of these online courses. Therefore, the MOOC providers can increase the effectiveness of their course to retain their enroled people. The main focus could be kept on to improve those parameters that are likely to reduce the attrition rates in MOOCs.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Adamopoulos, P. (2013). What makes a great MOOC? An interdisciplinary analysis of student retention in online courses.

Agarwal, S. (2013). Data mining: Data mining concepts and techniques. In *Machine Intelligence and Research Advancement (ICMIRA), 2013 International Conference on* (pp. 203–207). IEEE.

Al-Shabandar, R., Hussain, A., Laws, A., Keight, R., Lunn, J., & Radi, N. (2017). Machine Learning approaches to predict learning outcomes in Massive open online courses. In *Neural Networks (IJCNN), 2017 International Joint Conference on* (pp. 713–720). IEEE.

Arora, S., Goel, M., Sabitha, A. S., & Mehrotra, D. (2017). Learner groups in massive open online courses. *American Journal of Distance Education, 31*(2), 80–97.

Bassi, R., Daradoumis, T., Xhafa, F., Caballé, S., & Sula, A. (2014). Software agents in large scale open e-learning: a critical component for the future of massive online courses (MOOCs). In *Intelligent Networking and Collaborative Systems (INCoS),* 2014 *International Conference on* (pp. 184–188). IEEE.

Bates, T. (2013). Look back in anger? A review of online learning in 2013.

Bharara, S., Sabitha, S., & Bansal, A. (2017). Application of learning analytics using clustering data Mining for Students' disposition analysis. *Education and Information Technologies*, 1–28.

Bharara, S., Sabitha, S., & Bansal, A. (2018). Application of learning analytics using clustering data Mining for Students' disposition analysis. *Education and Information Technologies, 23*(2), 957–984.

Boyer, S., & Veeramachaneni, K. (2015). Transfer learning for predictive models in massive open online courses. In *International Conference on Artificial Intelligence in* Education (pp. 54–63). Springer, Cham.

Castro, E. G., & Tsuzuki, M. S. (2015). Churn prediction in online games using players' login records: A frequency analysis approach. *IEEE Transactions on Computational Intelligence and AI and Games, 7*(3), 255–265.

Chen, Y., Chen, Q., Zhao, M., Boyer, S., Veeramachaneni, K., & Qu, H. (2016). DropoutSeer: Visualizing learning patterns in Massive Open Online Courses for dropout reasoning and prediction. In *Visual Analytics Science and Technology (VAST), 2016 IEEE Conference on* (pp. 111–120). IEEE.

Dataverse. (2014). HarvardX-MITx Person-Course Academic Year 2013 De-identified Dataset, Version 2.0, https://dataverse.harvard.edu/file.xhtml?fileId=2468954&version=RELEASED&version=.0.

Gallén, R.C., & Caro, E.T. (2017). An exploratory analysis of why a person enrolls in a massive open online course within MOOCKnowledge data collection. In *Global Engineering Education Conference (EDUCON), 2017 IEEE* (pp. 1600–1605). IEEE.

Gamage, D., Fernando, S., & Perera, I. (2015) August. Factors leading to an effective MOOC from participiants perspective. In *Ubi-Media Computing (UMEDIA), 2015 8th International Conference on* (pp. 230–235). IEEE.

Hegyesi, F., Kártyás, G., & Gáti, J. (2017). Answers to the 21st century challenges at a university with technical training. In *Intelligent Systems and Informatics (SISY), 2017 IEEE 15th International Symposium on* (pp. 000365–000368). IEEE.

Huang, N.F., Hsu, H.H., Chen, S.C., Lee, C.A., Huang, Y.W., Ou, P.W., & Tzeng, J.W. (2017). VideoMark: A video-based learning analytic technique for MOOCs. In *Big Data Analysis (ICBDA), 2017 IEEE 2nd International Conference on*(pp. 753–757). IEEE.

Kaggle. (2017a). big_student_clear_third_version, https://www.kaggle.com/kanikanarang94/mooc-dataset/data.

Kaggle. (2017b). cs_mitx, MOOC Dataset, https://www.kaggle.com/chellaindu/mooc-dataset/data.

Kaveri, A., Gunasekar, S., Gupta, D., & Pratap, M. (2016). Decoding Engagement in MOOCs: An Indian Learner Perspective. In *Technology for Education (T4E), 2016 IEEE Eighth International Conference on* (pp. 100–105). IEEE.

Khalil, H., & Ebner, M. (2014). MOOCs completion rates and possible methods to improve retention-a literature review. In *EdMedia: World Conference on Educational Media and* Technology (pp. 1305–1313). Association for the Advancement of Computing in Education (AACE).

Kloft, M., Stiehler, F., Zheng, Z. & Pinkwart, N. (2014). 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* (pp. 60–65).

Liyanagunawardena, T.R., Parslow, P., & Williams, S. (2014). Dropout: MOOC participants' perspective.

Machado, N.L., & Ruiz, D.D. (2017). Customer: A novel customer churn prediction method based on mobile application usage. In *Wireless Communications and Mobile Computing Conference (IWCMC), 2017 13th International* (pp. 2146–2151). IEEE.

Mulik, S., Yajnik, N., & Godse, M. (2016). Determinants of acceptance of massive open online courses. In *Technology for Education (T4E), 2016 IEEE Eighth International Conference on* (pp. 124–127). IEEE.

Onah, D.F., Sinclair, J., & Boyatt, R. (2014). Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 proceedings* (pp. 5825–5834).

Rodriguez, C. O. (2012). MOOCs and the AI-Stanford like courses: Two successful and distinct course formats for massive open online courses. *European Journal of Open, Distance and E-Learning, 15*(2).

Rosé, C.P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P., & Sherer, J. (2014). Social factors that contribute to attrition in MOOCs. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 197–198). ACM.

Sabitha, A. S., Mehrotra, D., Bansal, A., & Sharma, B. K. (2016). A naive bayes approach for converging learning objects with open educational resources. *Education and Information Technologies, 21*(6), 1753–1767.

Sandanayake, T.C., & Madurapperuma, A.P. (2013). Computational model for affective e-Learning: Developing a model for recognising E-Learner's emotions. In *Innovation and Technology in Education (MITE), 2013 IEEE International Conference in MOOC* (pp. 174–179). IEEE.

Schaffer, J., Huynh, B., O'Donovan, J., Höllerer, T., Xia, Y., & Lin, S. (2016). An analysis of student behavior in two massive open online courses. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on* (pp. 380–385). IEEE.

Shah, D. (2015). By the numbers: MOOCs in 2015. *Class Central.*

Sharkey, M., & Sanders, R. (2014). A process for predicting MOOC attrition. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs* (pp. 50–54).

Shen, L., Wang, M., & Shen, R. (2009). Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment. *Journal of Educational Technology & Society, 12*(2), 176.

Shi, C., Fu, S., Chen, Q., & Qu, H. (2015). VisMOOC: Visualizing video clickstream data from massive open online courses. In *Visualization Symposium (PacificVis), 2015 IEEE Pacific* (pp. 159–166). IEEE.

Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning, 2*(1), 3–10.

Sooryanarayan, D.G., & Gupta, D. (2015). Impact of learner motivation on mooc preferences: Transfer vs. made moocs. In *Advances in Computing, Communications and Informatics (ICACCI),* 2015 *International Conference on* (pp. 929–934). IEEE.

Sunar, A., White, S., Abdullah, N., & Davis, H. (2016). How learners' interactions sustain engagement: a MOOC case study. *IEEE Transactions on Learning Technologies*.

Wu, Y., Pitipornvivat, N., Zhao, J., Yang, S., Huang, G., & Qu, H. (2016). Egoslider: visual analysis of egocentric network evolution. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 260–269.

Yousef, A.M.F., Chatti, M.A., Schroeder, U., & Wosnitza, M. (2014). What drives a successful MOOC? An empirical examination of criteria to assure design quality of MOOCs. In *Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on* (pp. 44–48). IEEE.

Zhou, N., Gifford, W.M., Yan, J., & Li, H. (2016). End-to-end solution with clustering method for attrition analysis. In *Services Computing (SCC), 2016 IEEE International Conference on* (pp. 363–370). IEEE.