A MACHINE LEARNING BASED APPROACH TO ENHANCE MOOC USERS' CLASSIFICATION

Dr. Youssef MOURDI
ORCID: 0000-0003-0999-6388
Faculty of Sciences Semlalia
Cadi Ayyad University
Marrakech, MOROCCO

Dr. Mohammed SADGAL
ORCID: 0000-0002-1087-0988
Faculty of Sciences Semlalia
Cadi Ayyad University
Marrakech, MOROCCO

Dr. Wafa BERRADA FATHI
ORCID: 0000-0002-6316-3208
Faculty of Sciences Semlalia
Cadi Ayyad University
Marrakech, MOROCCO

Dr. Hamada EL KABTANE
ORCID: 0000-0003-2864-1929
Faculty of Sciences Semlalia
Cadi Ayyad University
Marrakech, MOROCCO

Received: 17/02/2019 **Accepted:** 28/05/2019

ABSTRACT

At the beginning of the 2010 decade, the world of education and more specifically e-learning was revolutionized by the emergence of Massive Open Online Courses, better known by their acronym MOOC. Proposed more and more by universities and training centers around the world, MOOCs have become an undeniable asset for any student or person seeking to complete their initial training with free distance courses open to all areas. Despite the remarkable number of course enrollees, MOOCs have a huge dropout rate of up to 90%. This rate significantly affects the efforts made by the moderators for the success of this pedagogical model and negatively influences the learners' experience and their supervision. To address this problem and help instructors streamline their interventions, we present a solution to classify MOOC learners into three distinct classes. The approach proposed in this paper is based on the filters methods to select the most relevant attributes and ensembling methods of machine learning algorithms. This approach has been validated by four MOOC courses from Stanford University. In order to prove the performance of the model (92.2%), a comparative study between the proposed model and other algorithms was made on several performance measures.

Keywords: Distance Education, Dropout, Feature Selection, Educational Datamining, MOOC, Machine Learning.

INTRODUCTION

Since their creation in 2008 by Georges Siemens, Massive Open Online Courses (MOOCs) have revolutionized distance education by their quality and simplicity. They allow students and all those who wish to take free online courses in a variety of subject areas, to interact with other learners / professors even at

the other side of the world (ask questions , ask for help and support, etc.) (Sanchez-Gordon & Luján-Mora, 2016). If MOOCs are an advantage for students wishing to complete / improve their face-to-face training and for people who have a professional activity and wish to take courses at a suitable time without travel restrictions; they become a major necessity for universities whose infrastructure is no longer able to support the mass of students in perpetual growth. MOOCs are also used by training companies and private trainers who offer free or paid certifications via platforms such as Udemy, Cousera, Udacity and many others (Gupta & Sambyal, 2013).

For all these benefits and others, the MOOCs have generated a great deal of satisfaction among academics and a high demand to the point that the number of enrolled learners in courses is counted by hundreds of thousands, a number that hides, however, a very serious problem pointed and specific to MOOCs. This problem concerns the high drop-out rate since less than 10% of the enrollees complete the training (Liyanagunawardena, Parslow, & Williams, 2014). In other words, the impressive total enrolment rate coincides with a very low success rate. This has been proven by several feedbacks like a software engineering course offered by the University of MIT and Berkeley, which received 50,000 registrations but just 7% were able to pass the MOOC (Gupta & Sambyal, 2013). Another study (Onah, Sinclair, & Boyatt, 2014) cited the experience of Duke University that launched a Bioelectricity MOOC, a course that received 12175 registrations. Despite this huge number, only 7,761 learners (representing 64% of all learners) followed at least one video, 26% answered a quiz and only 2.6% completed the course.

Whatever the reasons behind this dropout rate, which is known to be extremely high (90%), it leads to considerable losses of the resources deployed by the teaching team and managers to make a MOOC course a successful experience. Also, this affects the quality of collective pedagogical activities such as projects by demotivating the rest of the group's learners. In addition, instructors and course facilitators are no longer able to master the task of coaching because they can no longer identify students at risk of leaving the MOOC. Determining precisely these learners, in addition to those who might succeed or fail, will effectively lead all efforts; and thereby streamlining the interventions made for each type of learner. In other words, a classification of learners into three distinct classes (class of learners passing the course, class of learners failing and those leaving the MOOC) is a first-class solution to the dropout problem.

This classification is made possible through the analysis of important data generated by MOOC platforms. This data collects the learners' personal information, their login and navigation data, their performance and their interactions in the forums and with the educational resources provided to them. Exploring these data effectively will extract interesting prediction and classification barometers through a relevant choice of features that model these data. This has led several researchers to follow this strategy, but according to the literature review, several works have focused solely on the development of solutions for the prediction of learners at risk of leaving the MOOC, the majority of whom have neglected the attributes selection stage.

Following all these motivations, this paper presents an approach based on feature selection methods and machine learning algorithms to automate the selection of the most relevant and effective characteristics for analysis and interpretation of the dropout phenomenon and to ensure the prediction and classification of learners in a MOOC. This approach begins with an analysis in collaboration with a set of pedagogues whose objective is to establish a set of initial characteristics that are important from a pedagogical point of view. Then after, a study was launched in order to choose the best method of selecting the characteristics and the best performing machine learning algorithm. In this study, four methods were used to select Filters family characteristics and six machine learning algorithms, one of which is the combination of several algorithms. Finally, an implementation was performed and tested on a dataset composed of four MOOC courses with 49,551 enrolled learners.

The main contributions of this research can be quoted as follows:

- To extract automatically the most significant features for the analysis, the classification and the prediction processes and to evaluate the use of filter-based feature selection methods.
- To look for the best combination of feature selection methods and machine learning algorithms for the best possible predictive model.
- To evaluate the performance and behaviour of an ensembling method based on the vote between several machine learning algorithms in the context of E-learning data and the prediction of learners at risk of quitting a MOOC course.

The present paper is organized into three main parts, the first presents a set of works that have been made in the same context and inspired our research. The second part presents the adopted methodology and all the materials used to carry out this research work. In the third and last part we display and discuss the results obtained.

RELATED WORKS

MOOCs are becoming more and more popular in the world of distance learning and are attracting a lot of interest from knowledge and certification researchers, which has resulted in the creation of a very large number of platforms around the world. According to (Feng, Tang, & Liu, 2019), these courses have attracted the attention of more than 81 million people who have enrolled in more than 9400 courses, significant numbers that continue to grow. These figures prove that MOOCs can guarantee an effective educational experience. A study conducted by Coursera in 2016, shows that MOOCs are very beneficial for learners who complete the courses, and especially in terms of their career, something that has been approved by more than 72% of learners in this survey (Chen et al., 2015).

On the other hand, the MOOCs with all their benefits, suffer from a problem that is very specific to them, it is the very large number of registered learners who abandon the course, a number that reaches 90% in some cases. This alarming finding has pushed researchers in the field of distance education to take this problem seriously, and as a result a lot of research has been launched by trying to propose solutions to address this dropout rate. This research has taken several paths, for example (Goel, Sabitha, & Choudhury, 2019) focused their study on understanding the environment of MOOCs and their problems, in particular the dropout problem. To carry out their research, the authors adopted the data mining techniques that allowed them to have the necessary factors and measures taken to push the similar types of learners to their maximum potential in the next MOOC courses. The authors put together a set of basic (initial) characteristics and then applied a chi-square test to determine which ones are related (correlated) to dropping out. This study is very interesting, but the number of initial characteristics taken was very small.

Many of the researches have focused on predicting student dropout by adopting machine learning techniques. For example, in (Halawa, Greene, & Mitchell, 2014), the authors present a predictor based on learners' activities for the prediction of those at risk of not completing the course. The proposed predictor analyses the activity of learners looking for signs of disability or interest that may cause learners to drop out or stay away for extended periods of time. For most courses, this model predicted between 40% and 50% dropout while learners were still active.

In another research (Chaplot, Rhim, & Kim, 2015), the authors propose an algorithm based on artificial neural networks to predict the attrition of students in MOOCs by using the sentiment analysis and show the significance of students' feelings in this task. As a result, the authors manage to ensure an accuracy of 72.1%. The authors relied on the forums discussions to classify learners, while according to (Xing, Chen, Stein, & Marcinkowski, 2016) and (Xing, Chen, Stein, & Marcinkowski, 2017) it is very clear that the learners' interactions in the discussion forums are very weak, which makes the effectiveness of the proposed contribution, a highly questionable model.

For the same objective, another research guided by learners' navigation traces and natural language pre-processing was initiated in (Crossley, Paquette, Dascalu, McNamara, & Baker, 2016). The purpose of this research was to examine whether students' online activity and the language they produce in the online discussion forum are predictive of success. This research was conducted on a sample of 320 learners who completed at least one rated task and produced at least 50 words in the discussion forums. The predictive model proposed guarantees an accuracy of 78%. Nevertheless, the sampling adopted in this research is very little for this model and its results to be generalized, given the nature of the MOOCs which are characterized in by their openings and therefore by the "massive" number of learners.

Using also Data Mining techniques, authors in (Burgos et al., 2017) analysed the archived data of the learners' notes to predict if a learner will drop a course. The authors deployed logistic regression models for classification purposes. In order to validate their proposal, the authors tested the proposed model on a set of 100 students. In parallel, the authors conducted a tutoring action plan to buy back the learners at risk. By adopting this approach, researchers are able to reduce the dropout rate by 14% compared to previous years.

In (Liu & Li, 2017), an explainable approach to find the reasons behind the dropout phenomenon convincingly using a data mining method in order to perform quantitative analyses. In this research, the authors tried to group learners using an unsupervised approach via the K-means algorithm and to determine the characteristics of learners who tend to abandon the MOOC. Subsequently, they analyzed the dropout factors in order to extract the reasons for these learners to leave their training.

To resume, the research inspired by these models has highlighted the complexity of the drpout phenomenon in MOOCs and developed prediction approaches based on recent techniques such as machine learning or data mining. However, these methods have focused on a limited number of prediction features that remain relatively similar.

METHODOLOGY AND MATERIALS

In this section, we present the methodology used to carry out this research work and the different tools that contributed to the development of the framework.

Research Process Overview

The proposed approach is divided into several phases. The first phase is features engineering, that unfortunately does not get much investment in the majority of the work that was done in the same context. It makes it possible to establish a set of basic characteristics which are significantly related to the field studied. Consequently, this step was conducted in contribution with some experts. The second step of our approach is data pre-processing, in which the extraction, reconstruction and necessary transformations of the data are carried out taking into account the characteristics set by the experts. Once the data is extracted, it is subsequently cleaned up to eliminate any unnecessary information stored in a data warehouse. Towards the end of the pre-processing stage, the data are standardized to unify the units of measurement and thus make possible the comparison between the data. Figure 1 shows the different steps followed to carry out this process.

In the third phase, we took four methods of features selection in order to reduce the dimensionality of the characteristics posed by the experts and to take only the most relevant ones. The characteristics returned by each feature selection method are subject to performance testing of the various automatic learning algorithms.

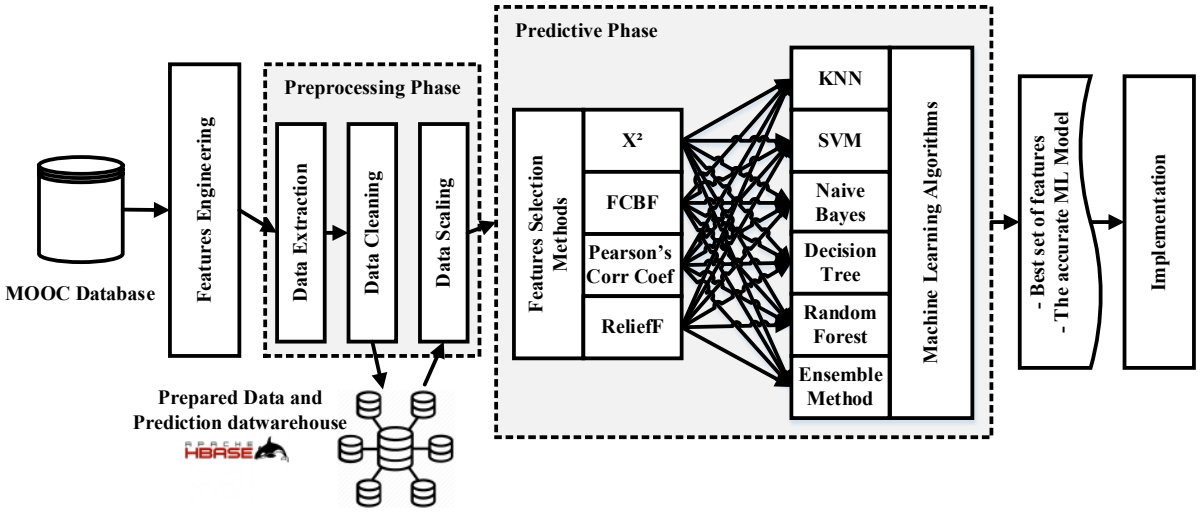


Figure 1. Research process overview

In what follows, we detail each phase, the tools used, the structure of the database and the way in which the prediction is ensured

Stanford Dataset

The dataset adopted in this research consists of four courses taught by Stanford University. The first course was on quantum mechanics, this course was divided into two parts, a part was assured during the year 2016 and its second part was proposed in 2017. For the second course, it is a course of algorithmic that was launched in two parts. These courses received a total of 49,551 registered learners, which makes this dataset a good sampling of testing and validation of our aproach.

The data is anonymous and divided into several Comma-Separated Values (CSV) files extracted from the OpenEdx platform. The table 1 shows each file used by presenting a brief description of its contents. For more information on these files, we recommend visiting the CAROL Stanford website (https://datastage.stanford.edu/).

Table 1. Dataset files and content

CSV File	Content
Demographics	Contains learner demographic information such as gender, year of birth and academic level. These information may be empty or null.
EventExtract	Contains all the information concerning the users' navigation on the platform. This data includes interaction with videos, transcripts, forum discussions and issues sent.
ActivityGrade	Includes learner score data in quizzes, and includes good and bad answers, answers validated by each learner, date of first submission, and date and time of last submission.
Forum	Contains threads of the speakers in the forum.
allData	Includes characteristics representing commitment as the number of connections, and the number of events in each session.
weeklyEffort	Contains the effort provided by each learner in a week (in seconds)

Features Engineering Phase and Extraction Phase

Before discussing the technologies used in this phase, we begin by exposing the adopted characteristics that guide the export of the data. Generally, the raw data in the MOOC platform databases is not directly exploitable and therefore feature engineering remains an essential phase. This will allow for the selection of initial indicators and predictors. In order to build all these basic characteristics, human expertise has been used, working with six teachers who already have at least one experience in a MOOC animation.

Also referring to the literature, 61 characteristics were retained and grouped under 11 categories that encompass not only the learners' navigation information on the MOOC platforms but also the interaction with the videos and their transcripts, the learner performance and effort provided during each week, the personal information, the navigation information, and the interactivity of learners with each other (Forum), prerequisites and use of additional resources. The table 2 presents the characteristics retained.

Table 2. The retained features and their categories

Category of features	Features
Video interaction	Number of completed videos in chapter 1, 2, 3, 4, 5, 6, 7, 8 and 9
	Number of times the learner tries to go back in the videos
	Number of times the learner tries to move forward in the videos
	Number of times the learner tries to speed up the video
	Number of times the learner tries to speed down the video
	Number of times the learner tries to play videos
	Number of times the learner ties to pause videos
	Number of times the learner ties to stop videos
Transcript interaction	Number of the transcript's downloads
	Number of times the learner interact with the transcripts
Quiz interaction	Number of sent quizzes
	Number of quizzes whose score is greater than 50% of the score defined by the instructor
	Number of quizzes whose score is less than 50% of the score defined by the instructor
	Number of attempts to send quizzes
	Number of 100% correct quizzes
	Average time between two quizzes sent (in minutes)
Effort	Time spent on the platform
	Number of connections
	Average number of days between two connections
	Number of active days of which the learner was logged on the platform
Personal information	Academic level
	Age
	Gender
Forum	Number of learner's thread response
	Number of created threads
	Number of up votes
	Number of down votes
Performance	Weekly final grade
Navigation	Number of views of course information
	Number of forum access
	Number of visits to the progress page
	Number of accessed chapters
	Number of visited sequential
	Number of reference access
Weekly Final Test	Number of answer in the weekly final Test in week 1, 2, 3, 4, 5, 6, 7, 8 and 9
Prerequisites	Number of completed prerequisites videos
Supplementary resources	Number of week 1, 2, 3, 4, 5, 6, 7, 8 and 9 supplementary resources access

Speaking about the used tools to extract and transform data, this module was developed based on Apache Spark which is a very sophisticated way for the processing of massive data. This choice was made by referring to the nature of our dataset that is distributed and massive, things that make search and access to information a very expensive task in terms of processor and RAM.

Spark offers a SparkSQL module that is quite comprehensive and offers a package of features to launch SQL queries and ensure joins between separate data (Meng et al., 2015) (Armbrust et al., 2015). Therefore, SparkSql has been an added value for data mining and construction.

Preprocessing Phase

After the extraction phase, the generated data has undergone two necessary operations. First, the data cleaning that consisted of detecting and eliminating incomplete observations. The second operation was the standardization of this data. This phase is very important because in the dataset generated, there are information with different scales (example: age in years, time spent on the platform in seconds, the time between two connections in days, etc.). The standardization of the data made it possible to adjust these values to make them comparable. For this, we used the MinMax method. In MinMax, the values of the entities are [0, 1] scaled as follows:

$$x_i^{new} = \frac{x_i - \min(X)}{\max(X) - \min(X)} \quad (1)$$

Where X is a relevant characteristic, is a possible value of X in the dataset and the normalized value .

The dataset also contains different characteristics: quantitative characteristics that cause no problem and qualitative data (nominal with more than 2 modalities) that must be transformed into numerical data so that they can be used during the learning phase of the algorithms. For this purpose, all nominal variables were transformed into fictitious data.

Prepared Data and Prediction Storage

For the storage of the prepared data, we used a structured model that combines in a single large column-oriented table, all the necessary tables. This model eliminates joins and thus allows a more optimal computation time. In this paper, the Apache Hadoop HBase distributed data management system was used (Vora, 2011).

HBase manages data within large tables (HTable) composed of rows (Row) and families of columns (Family). These are subdivided into columns (Qualifier). Lines are unique value identifiers (White, 2012). Following the same principle, the database created with HBase is modeled by a snowflake diagram (see Figure 2), where the dimensions are organized in a hierarchy. Each member belongs to a particular hierarchical level (or level of granularity).

The main table is the Learner_Class fact table. It contains the ubiquitous information (measures) in a learner's classification (class and score) as well as measures appropriate to the characteristics previously selected by the experts. This fact table is linked to a set of dimensions using foreign keys. First, the time dimension is spread over the week as the predictions are made weekly.

The dimension "Course" identified by a unique identifier of a course is spread over 3 hierarchical levels starting from the global course where information on a course are stored. The course consists of a set of "Chapter". These chapters are composed of several "Sequential" which themselves collect a number of "Video". Finally, the Learner dimension records a learner's general information such as academic level, age, and gender.

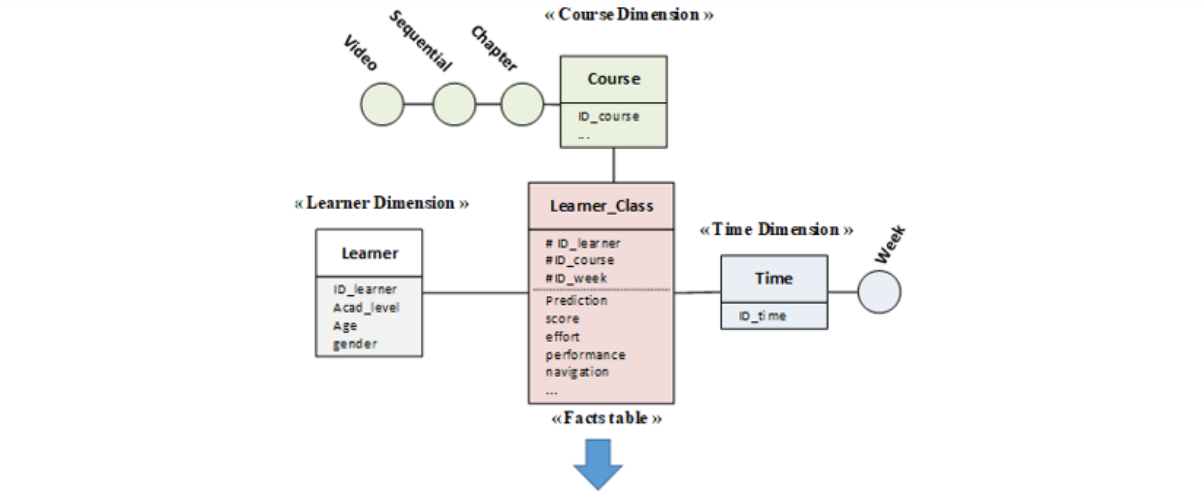


Table (Learner_class)	Row key	Learner_class		Learner Column Family		Course Column Family								Time Column Family		Timestamp
						Course		Chapter		Sequential		Video		Week		
	ID_learner + ID_course + ID_time + timestamp	prediction	...	age	...	name	...	Number_videos	...	title_sequential	...	duration	...			

Figure 2. HBase stucture

PREDICTIVE MODULE

Features Selection and Dimantionality Reduction Phase

In general, the adoption of datasets with a huge number of characteristics during the learning and testing phase of machine learning algorithms causes several problems that are remarkably detrimental to the performance of the predictors and, above all, lead to over-fitting models (Talavera, 2005)(Salcedo-Sanz, Cornejo-Bueno, Prieto, Paredes, & Garcia-Herrera, 2018).

In order to overcome the problems mentioned above, reducing the dimensionality of datasets is one of the most powerful tools. This power lies in the selection of the richest subset of characteristics terms of information (Alonso-betanzos, 2007). According to (Li et al., 2018), having a dataset with significant characteristics allows to:

- Remarkably improve the predictive performance of a machine learning model.
- Decrease the complexity of the model.
- Gain in terms of computing cost and resources.
- Avoid over-adjusting algorithms.

Although experts in the field can eliminate few irrelevant attributes, selecting the best subset of features usually requires a systematic approach. Currently, there are three families of automatic feature selection methods namely: filters (Talavera, 2005) (Alonso-betanzos, 2007), wrappers (Karegowda, Manjunath, & Jayaram, 2010) and embedded (Jovic, Brkic, & Bogunovic, 2015).

In this research, we focused on selection methods based on Filters. For this, we made use of four different methods which are listed below.

a. Chi-Square (X^2)

In accordance with (Bahassine, Madani, Al-sarem, & Kissi, 2018), the chi-square test is used in statistics to test the independence of two events. With the dataset on two events, we can get the observed account "O" and the expected account "E". Chi's squared score measures the difference between the expected count "E" and the count "O" observed.

In the feature selection, both events are an occurrence of the characteristic and an instance of the class. In other words, we want to test whether the occurrence of a specific feature and the occurrence of a specific class are independent. If both events are dependent, we can use the occurrence of the entity to predict the occurrence of the class. Our objective is to select the characteristics whose occurrence depends strongly on the occurrence of the class.

When the two events are independent, the number observed is close to the expected number, so it is a small chi-square score. So a high value of Chi-Square indicates that the independence assumption is incorrect. In other words, the higher the Chi-Square's score, the more likely the functionality is to be correlated to the class, so it must be selected for model learning.

b. Fast Correlation-Based Filter (FCBF)

According to (Khourdifi & Bahaj, 2018) and (Jain, Jain, & Jain, 2018), the Fast Correlation-Based Filter (FCBF) algorithm consists of two steps: the first one is a relevance analysis, aimed at classifying the input variables according to a relevance score, calculated as a symmetric uncertainty with respect to the output target. This step is also used to ignore irrelevant variables, which are those whose ranking score is below a predefined threshold. The second step is a redundancy analysis to select the predominant features in the relevant set obtained in the first step. This selection is an iterative process that removes variables that form an approximate Markov coverage.

c. Relief

It is an algorithm developed by Kira and Rendell in 1992 that uses a filtering method to select entities that are particularly sensitive to interactions between them. It was originally designed for binary classification problems with discrete or numerical characteristics. Relief calculates a score for each attribute, which can then be applied to rank and select the best performing ones. These scores can also be applied as a feature weight to guide downstream modelling. The notation of the features in relief is based on the identification of the differences of value of the characteristics between the pairs of nearest instances. If a feature value difference is observed in a pair of neighbouring instances with the same class, the feature score decreases. Alternatively, if a feature value difference is observed in a pair of neighbouring instances with different class values, the feature score increases (Urbanowicz, Meeker, Lacava, Olson, & Jason, 2018).

d. Pearson's correlation coefficient

The principle of this method is based on the computation of correlation between two variables x and y. The returned measure is exactly 0 if the two variables x and y are independent. In the case where the two variables are dependent, the measure is thus a value in the interval -1 and +1 indicating the dependence and the level, one says in this case that the two variables of which negatively or positively correlated. This method is commonly used to estimate the magnitude of the association between characteristics and class for a dataset (Mu, Liu, & Wang, 2017) (Ly, Marsman, & Wagenmakers, 2018).

Machine Learnig Algorithms

With regard to the learning algorithms used, the choice of five of them was based on their wide use in the literature. These algorithms are as follows:

- Support Vector Machine (SVM) (Naghibi, Ahmadi, & Daneshi, 2017)
- K Nearest Neighbors (KNN) (Martinez-Espana et al., 2018)
- Decision Tree (DT) (Erel, Stern, Tan, & Weisbach, 2018)
- Naive Bayes (NB) (Gao, Cheng, He, Susilo, & Li, 2018)
- Logistic Regression (LR) (Os, Ramos, Hilbert, & Leeuwen, 2018)

Our study was not limited to classical machine learning algorithms but also included the combinatorial method. According to the literature, the work proposed to predict a learner's class was based on the use of a single algorithm among the classical machine learning algorithms. On the other hand, one can of course improve the performances of prediction accuracies by combining several algorithms, which has been proven in several studies (Nagi & Bhattacharyya, 2013) (Sikora & Al-Laymoun, 2014). This principle is known as "ensembling methods" which group together several families namely Boosting (Sikora & Al-Laymoun, 2014) (Zhu, Xie, Wang, & Yan, 2017), Bagging (Choudhury & Bhowal, 2015) (Kabir, Ruiz, & Alvarez, 2014) and combinatorial methods. The first two classes of methods work with a single algorithm called "weak" to generate a stronger model. While combinatorial methods combine several algorithms at once in order to have a more powerful predictive model (Nagi & Bhattacharyya, 2013) (Zitlau et al., 2016) (Alves, 2017).

Depending on the nature of prediction (classification or regression), we find combinatorial methods that are based on the vote or the average of the predicted values. All algorithms share the same set of learning data. During the test phase, each algorithm autonomously makes its decision, then after all decisions are transmitted to a voting or averaging module to have the final decision. In this research and since we are dealing with a problem of classification, we have adopted a combinatorial method based on voting as shown in Figure 3.

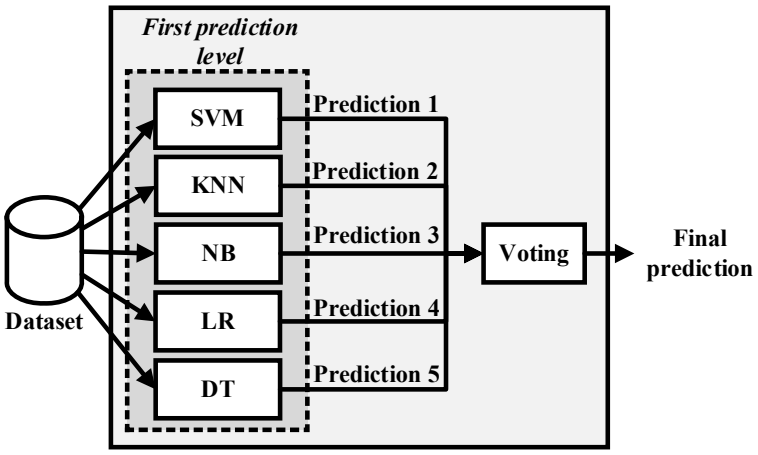


Figure 3. Ensemble model based on voting

In order to compare and decide on the feature selection method and the machine learning algorithm to adopt, several performance indicators (table 3) were set. The study of the machine learning algorithms performances was done by creating, at each time, a model with the features returned by a selection method.

Table 3. Performance mesures and accuracy

Measure	Formula	
Accuracy	$$\text{Accuracy} = \frac{\text{TruePositives} + \text{FalseNegatives}}{\text{Total Number of Sample}}$$	(2)
Precision	$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}}$$	(3)
AUC	The ROC curve represents the true positive rate (TPR) based on the false positive rate (FPR), with:	
	$$\text{TVP} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}}$$	(4)
	$$\text{TFP} = \frac{\text{FalsePositives}}{\text{FalsePositives} + \text{TrueNegatives}}$$	(5)
	The AUC is the area below the ROC curve and is calculated as:	
	$$\text{AUC} = \int_0^1 \text{ROC(t)}.\text{dt}$$	(6)

With:

- TruePositives: The cases in which we predicted YES and the actual output was also YES.
- TrueNegatives: The cases in which we predicted NO and the actual output was YES.
- FalsePositives: The cases in which we predicted YES and the actual output was NO.
- FalseNegatives: The cases in which we predicted NO and the actual output was also NO.

As it is known in the world of machine learning, and in order to generate a predictive model, it is necessary to generally go through two phases, the first is the learning phase allowing the algorithm to be situated with respect to the learning data. The second step being the generalization of the model, which makes it possible to test the predictive performances of the model on data never seen. In the same context, we divided the data into two distinct parts and by courses. In other words, we adopted three courses for the learning of the models and the data of a single course for the test (70% of the data for learning and 30% of the data for the test).

RESULTS AND DISCUSSION

In this part of the paper, we detail the results obtained after the implementation of the proposed framework. For this purpose, we first present the results obtained by combining automatic learning algorithms and the different methods of selecting characteristics. In a second place, we devote the second part to the discussion where we show some user interfaces of the implemented predictive module

Features Selection Methods Comparison and Predictive Models Performance

This section illustrates the performance (accuracy, precision, and AUC) of machine learning algorithms with and without the feature selection methods presented previously. These performances were evaluated over the 9 weeks of the course, but the results presented in this section are for only three weeks, namely weeks 3, 5 and 7.

Without use of any Selection Method

this part, we present the performance results of the different machine learning algorithms taking into account all the features, in other words, without the adoption of any selection method. Referring to Table 4, which shows the machine learning algorithms performances over 3 weeks of the course (test dataset), we can clearly see that the predictive performance of the combinatorial model based on the vote in terms of accuracy are quite interesting compared to the rest of the models. Testing the precision of the different algorithms on all the dataset attributes will allow us to evaluate the importance and usefulness of the integration of the features selection methods.

Table 4. Machine algorithms performance measures over 3 weeks without features selection methods

Week	Algorithm	ACCURACY	AUC	PRECISON
3	SVM	0,844	0.831	0.811
	KNN	0,81	0.788	0.803
	DT	0,788	0.791	0.775
	NB	0,835	0.822	0.801
	LR	0,853	0.840	0.823
	Voting	0,882	0.892	0.891
5	SVM	0,801	0.872	0.888
	KNN	0,781	0.815	0.859
	DT	0,764	0.800	0.871
	NB	0,832	0.845	0.822
	LR	0,856	0,869	0,874
	Voting	0,88	0.890	0.889
7	SVM	0,865	0.849	0.861
	KNN	0,822	0.809	0.857
	DT	0,76	0.809	0.884
	NB	0,819	0.836	0.851
	LR	0,868	0,852	0,860
	Voting	0,878	0.869	0.837

Using the X^2 Feature Selection Method

The first method evaluated is the X^2 method (table 5), for which the results obtained by the machine learning algorithms (over weeks 3, 5 and 7) are presented. The results show that with this method, LR is the most efficient among the classical algorithms. Also, it must be pointed out that the voting-based ensembling method generates more interesting performances than the rest of the algorithms.

Table 5. Machine algorithms performance measures over 3 weeks using X^2 method

Week	Algorithm	ACCURACY	AUC	PRECISON
3	SVM	0,857	0.859	0.881
	KNN	0,803	0.789	0.809
	DT	0,802	0.813	0.782
	NB	0,826	0.868	0.881
	LR	0,86	0,872	0,875
	Voting	0,884	0.903	0.918
5	SVM	0,821	0.864	0.874
	KNN	0,812	0.826	0.861
	DT	0,815	0.829	0.865
	NB	0,846	0.872	0.849
	LR	0,834	0,852	0,863
	Voting	0,884	0.908	0.909
7	SVM	0,874	0.876	0.874
	KNN	0,861	0.829	0.872
	DT	0,742	0.871	0.852
	NB	0,773	0.860	0.871
	LR	0,872	0,889	0,880
	Voting	0,875	0.906	0.918

Using the FCBF Feature Selection Method

Concerning the FCBF selection method, we conclude that the results do not change because LR remains the most accurate algorithm with a greater accuracy value than that obtained by the X^2 method. This is also true for the vote-based ensembling model as shown in Table 6.

Table 6. Machine algorithms performance measures over 3 weeks using FCBF method

Week	Algorithm	ACCURACY	AUC	PRECISON
3	SVM	0,851	0.861	0.876
	KNN	0,816	0.792	0.826
	DT	0,822	0.824	0.789
	NB	0,834	0.872	0.897
	LR	0,849	0,878	0,900
	Voting	0,889	0.918	0.922
5	SVM	0,842	0.848	0.840
	KNN	0,829	0.853	0.861
	DT	0,819	0.833	0.863
	NB	0,853	0.877	0.852
	LR	0,836	0,905	0,872
	Voting	0,913	0.929	0.926
7	SVM	0,878	0.880	0.904
	KNN	0,857	0.767	0.870
	DT	0,744	0.870	0.855
	NB	0,839	0.875	0.875
	LR	0,893	0,868	0,900
	Voting	0,903	0.914	0.921

Using the Relief feature selection method

The table 7 clearly shows that the SVM exceeds the other algorithms in terms of performance, by adopting the Relief selection method. However, its performance does not go beyond the voting ensembling model

Table 7. Machine algorithms performance measures over 3 weeks using Relief method

Week	Algorithm	ACCURACY	AUC	PRECISON
3	SVM	0,857	0.873	0.879
	KNN	0,835	0.824	0.851
	DT	0,809	0.839	0.829
	NB	0,832	0.864	0.899
	LR	0,862	0,912	0,893
	Voting	0,889	0.937	0.937
5	SVM	0,895	0.888	0.906
	KNN	0,832	0.843	0.831
	DT	0,83	0.821	0.912
	NB	0,86	0.868	0.877
	LR	0,842	0,845	0,914
	Voting	0,937	0.958	0.933
7	SVM	0,898	0.881	0.876
	KNN	0,854	0.879	0.875
	DT	0,752	0.874	0.849
	NB	0,866	0.875	0.880
	LR	0,897	0,899	0,927
	Voting	0,929	0.928	0.947

Using the Pearson's Correlation Coefficient Method

Finally, being combined with the Pearson's Correlation Coefficient method, SVM outstrips the other algorithms by generating better performances over the three weeks. But we conclude that the predictive model based on the vote ensures the best accuracy over the three weeks studied. The table 8 shows the results returned using this features selection method.

Table 8. Machine algorithms performance measures over 3 weeks using the Pearson's correlation coefficient method

Week	Algorithm	ACCURACY	AUC	PRECISON
3	SVM	0,844	0.862	0.880
	KNN	0,822	0.854	0.863
	DT	0,818	0.830	0.847
	NB	0,84	0.874	0.905
	LR	0,866	0,899	0,928
	Voting	0,887	0.941	0.947
5	SVM	0,89	0.898	0.912
	KNN	0,847	0.858	0.863
	DT	0,858	0.832	0.899
	NB	0,918	0.888	0.897
	LR	0,854	0,935	0,909
	Voting	0,939	0.950	0.927
7	SVM	0,898	0.892	0.886
	KNN	0,85	0.868	0.875
	DT	0,732	0.884	0.852
	NB	0,869	0.875	0.890
	LR	0,89	0,900	0,898
	Voting	0,969	0.958	0.948

DISCUSSION

In this research, we aimed to propose a predictive model in order to classify students in MOOC courses into three classes, namely learners at risk of leaving the MOOC, those likely to fail and finally those with a high chance of succeeding. For this, a study was conducted to find the best combination of filter selection methods and a set of algorithms and machine learning techniques giving the most accurate predictions.

In the previous section, the performance results obtained for each algorithm combined with each selection method were spread over 3 weeks of the course. In this second part, we use these results on average of accuracies over the 9 weeks of the test course.

First, the table 9 presents the average of the weekly predictions without and with recourse to the 4 methods of features selection adopted in this work. First, we discuss the precision of the algorithms with all the characteristics of the dataset, in this case without the use of any selection method. We therefore note that DT is the least efficient algorithm followed by KNN with respective average values of 77% and 83.9%. A slight difference of 0.6% is recorded between the performances of SVM and NB which are exceeded by LR, the strongest algorithm among the classical algorithms. Finally, an average prediction of 92% is provided by the voting-based ensembling model previously described in this work.

Secondly, we note that the DT algorithm remains the weakest (never exceeding 80% accuracy) on all filters-based selection methods. On the other hand, the other algorithms perform differently according to the selection method adopted. For example, SVM gives the best accuracy with the X^2 method but a less interesting performance with the FCBF method. KNN and NB, on the other hand, generate the best performances when combined with the Relief method with average values of 83.4% and 83.6% respectively. Unlike DT, the vote-based model remains the best regardless the method adopted.

Table 9. Machine algorithms accuracy average with and without the use of feature selection methods

Machine learning algorithm	Without	χ^2	FCBF	Relief	Pearson
SVM	0,852	0,849	0,834	0,843	0,839
KNN	0,839	0,828	0,818	0,834	0,82
DT	0,77	0,776	0,762	0,776	0,769
NB	0,858	0,815	0,836	0,852	0,848
LR	0,868	0,855	0,871	0,854	0,857
Voting	0,92	0,882	0,901	0,922	0,907

The features selection phase importance marked in section (Features selection and dimantionality reduction phase) is evidenced by the results obtained during this research. By adopting a feature selection method, one can notice on the graph of the figure 4, that almost all the algorithms give average prediction values relatively close, or even more important to those obtained in the case where no dimensionality reduction has been integrated. That said, going through a selection phase reduces the complexity of a machine learning model and therefore avoids all the problems mentioned above in section (Features selection and dimantionality reduction phase).

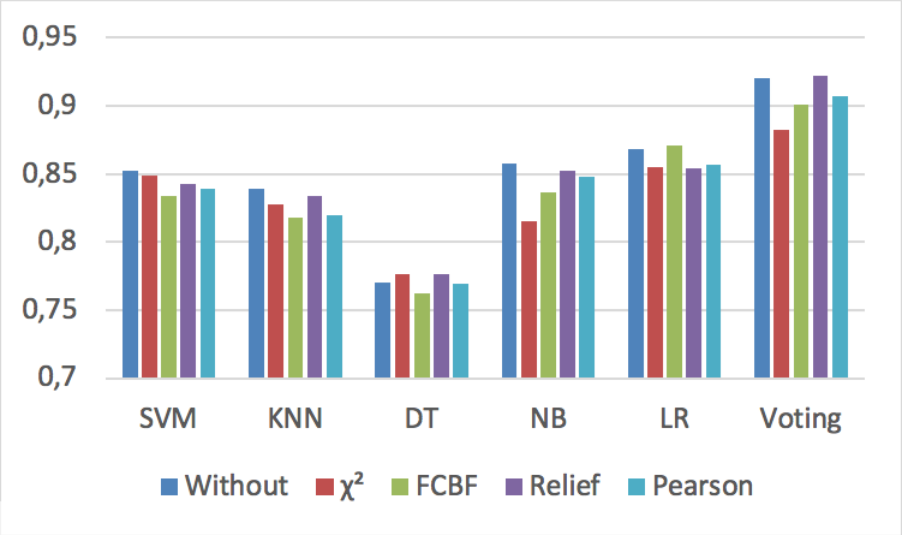


Figure 4. The average accuracy of the machine learning algorithms combined with the different features selection methods

Taking into consideration the results obtained in the comparative study, we opted for the implementation of a predictive module based on both the method of selection of features Relief and the combinatory method 'Voting'. The latter will give a more precise and refined vision to the instructors concerning the different classes of learners in a MOOC in a weekly frequency.

Figures (figure 5 and 6) give an idea of what the framework will generate as available interfaces to instructors. In figure 5, instructors can choose a course from the list of courses they are responsible for in addition to the week for which they wish to view the predictions of the learner classes.

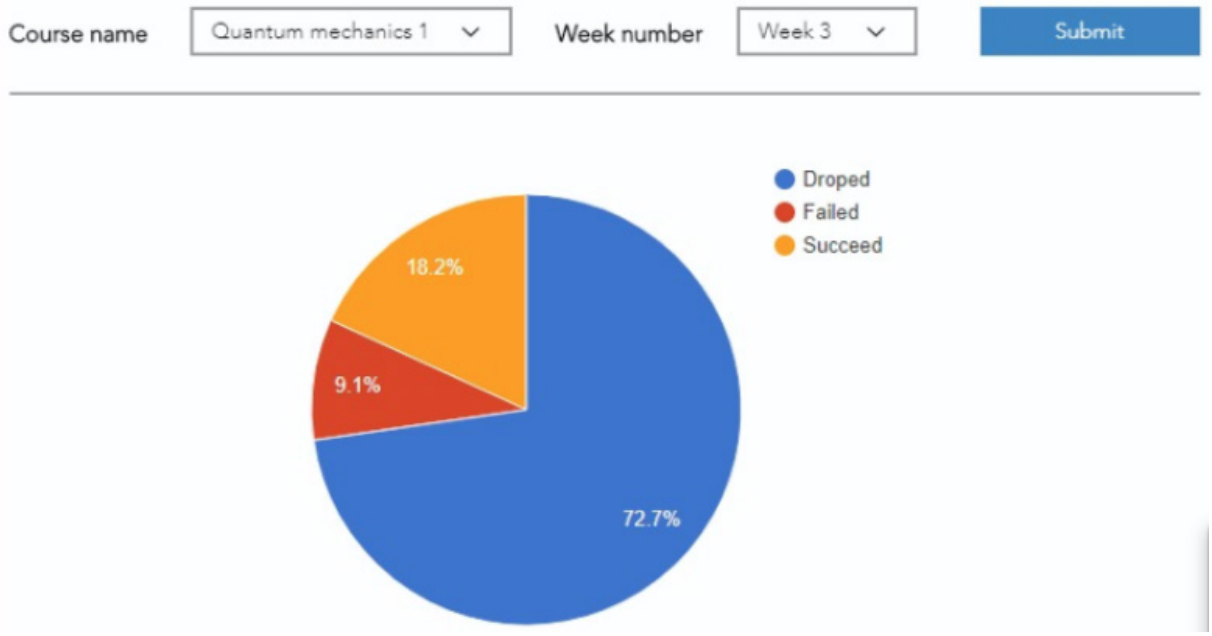
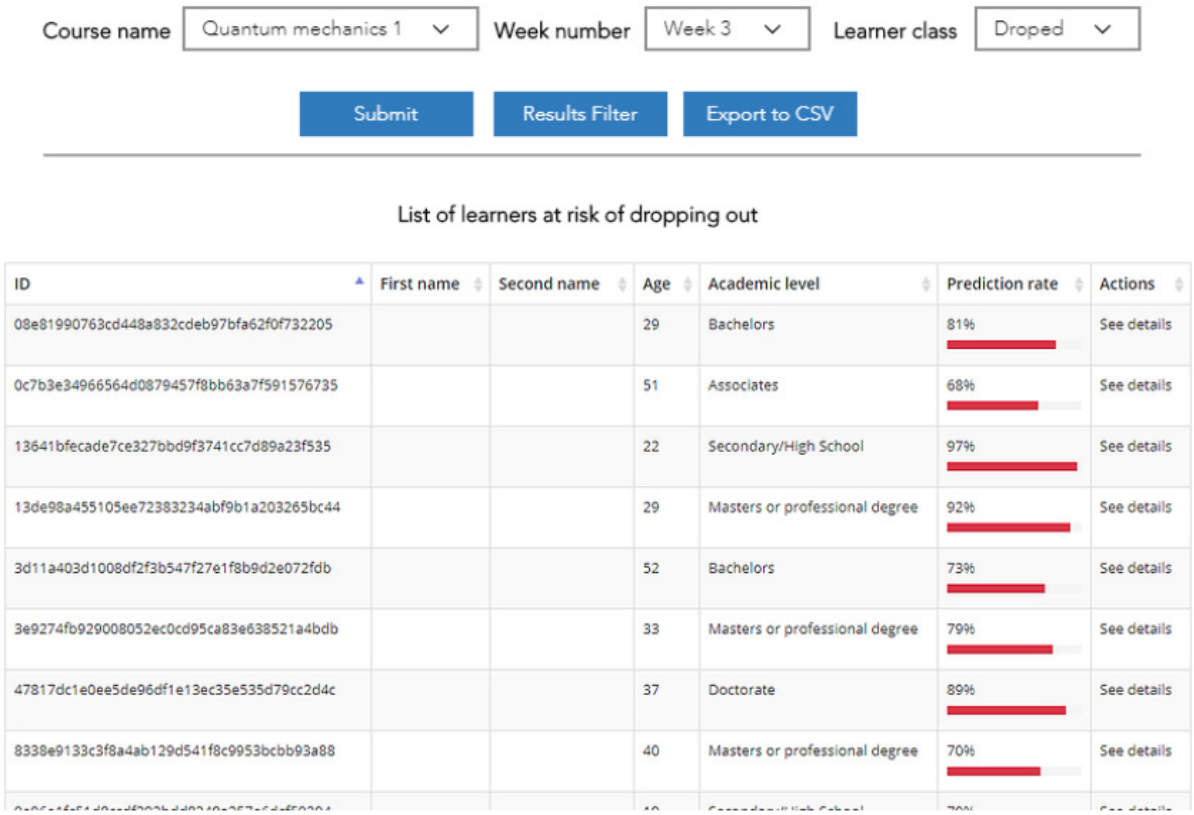


Figure 5. Distribution of learners after classification using the predictive framework

For more details, an instructor can select a class of learners to view the list of learners belonging to it with the score returned by the predictive module of each learner. Figure 6 gives an example of the droped class.



ID	First name	Second name	Age	Academic level	Prediction rate	Actions
08e81990763cd448a832cdeb97bfa62f0f732205			29	Bachelors	81%	See details
0c7b3e34966564d0879457f8bb63a7f591576735			51	Associates	68%	See details
13641bfecade7ce327bbd9f3741cc7d89a23f535			22	Secondary/High School	97%	See details
13de98a455105ee72383234abf9b1a203265bc44			29	Masters or professional degree	92%	See details
3d11a403d1008df2f3b547f27e1f8b9d2e072fdb			52	Bachelors	73%	See details
3e9274fb929008052ec0cd95ca83e638521a4bdb			33	Masters or professional degree	79%	See details
47817dc1e0ee5de96df1e13ec35e535d79cc2d4c			37	Doctorate	89%	See details
8338e9133c3f8a4ab129d541f8c9953bcbb93a88			40	Masters or professional degree	70%	See details

Figure 6. List of learners returned by the predictive framework with their classes and the prediction rate

The module will browse the learner database one by one and offers, to the instructors, a list of the learners with their classes and prediction scores (the probability that the learner belongs to the predicted class). At this point, it is up to the instructors to make the appropriate decision and intervene in the current week or wait for the next week's predictions. In other words, a learner who was classified as "at risk of dropping out" with a score of 80% requires an urgent intervention from the instructors unlike another who was classified in the same class with a score of 50 % or less. In this case, instructors cannot make a decision and must wait for the next week's predictions.

CONCLUSION AND FUTURE WORKS

Online learning or E-learning, has been developed in different forms with the emergence of Internet technologies and communication, among these forms the Massive Open Online Courses. Known to be accessible to all and often free, MOOCs suffer from a huge dropout rate that reaches 90% in some courses. Several researchers have therefore been interested in the reasons behind this large number, but many others have seen the need to predict learners at risk of dropping out.

Contrary to several research works, the approach proposed in this paper allows the detection not only of the learners at risk of leaving the MOOC but also generates weekly predictions to determine also the learners who are towards the path of the success and, therefore obtaining their certifications and those likely to fail. These predictions and the resulting indicators are based on the evolving field of artificial intelligence, including machine learning tools. Combined with the rise of Big Data, machine learning algorithms can automate actions through fast and efficient data analysis.

The proposed approach is based on a set of phases, the important ones are the features selection and the classification. For the first stage, the selection of the most relevant characteristics, a step often neglected in several solutions; was realized thanks to different filters methods which showed high performances. The predictive module, meanwhile, is a module based on voting grouping methods. This module, evaluated under various performance measures, provides weekly forecasts in a MOOC with an average accuracy of

92.2%. The results returned by our model are very promising and far exceed the models of the literature in terms of accuracy of prediction and performance.

The objective of proposing this approach is to give the MOOC instructors and trainers, through an interactive user interface, the opportunity to ensure rational, targeted and effective interventions for each class of learners. These interventions can be offered in the form of support courses, additional resources or any other assistance aimed specifically at the classes of learners at risk. But for more successful interventions, we will, in a future work, identify the causes of MOOC dropout while looking for a way to automate the intervention to retain this class of learners (at risk of abandonment or failure).

Acknowledgements: This research was done through Stanford University's Advanced Research Center on Online Learning (CAROL), which we thank immensely for all the facilities they provided for us. We also wish to express our full gratitude to Ms. Kathy Mirzaei for her responsiveness as well as her collaboration. We wish to warmly thank Mr. Mitchell Stevens, Director of Digital Research and Planning, as well as all the CAROL commission for the trust they have given us.

BIODATA and CONTACT ADDRESSES of AUTHORS

Dr. Youssef MOURDI, is a doctor of computer science from cadi ayyad university, he obtained his master's degree in information systems engineering from the same university in 2013. Currently he is interested in the field of intelligent education by integrating more and more the field of automatic learning in the field of distance education

Youssef MOURDI
Computer Science, Faculty of sciences SEMLALIA
Address: CADI AYYAD University, 40 000, MARRAKECH, MOROCCO
Phone: +212 667 835 667,
E-mail: mourdiyoussef@gmail.com

Dr. Mohammed SADGAL, is a professor of computer science at Cadi Ayyad University (Marrakech) and he is doing research on computer vision with the Vision team at the LISI Laboratory. His research interests include object recognition, image understanding, video analysis, multi-agent architectures for vision systems, 3D modeling, virtual and augmented reality, among other topics (see the research link for details). Before Marrakech, He was in Lyon (France), working as Engineer in different computer Departments between 1988-1994. He obtained a PhD in 1989 from Claude Bernard University (France, Lyon).

Mohammed SADGAL
Computer Science, Faculty of sciences SEMLALIA
Address: CADI AYYAD University, 40 000, MARRAKECH, MOROCCO
Phone: +212 679 850 461
E-mail: sadgal@hotmail.com

Dr. Wafa BERRADA FATHI, is a PhD in computer graphics systems, at the University of Strasbourg, in 1988. She is an E-learning Expert with Master's degree «UTICEF»: (use of information technologies and communication in the education and the formation), at the University of Strasbourg, in 2009. Its research interest focus on integrating virtual reality especially 3D virtual world in the teaching methodology and follow its impact on student motivation and performance.

Wafa BERRADA FATHI
Computer Science, Faculty of sciences SEMLALIA
Address: CADI AYYAD University, 40 000, MARRAKECH, MOROCCO
Phone: +212 661 179 359,
E-mail: w.f.berrada@gmail.com

Dr. Hamada EL KABTANE, is received the M.S. degree from the Faculty of Sciences Ibn Tofail of Kenitra, Morocco, in 2012 and then he holds a PhD degree in the Information Systems Engineering Laboratory (LISI) in the Faculty of Sciences, Cadi Ayyad University of Marrakech, Morocco. He worked on the Virtual Learning Environments, the 3D Image Processing, Virtual Reality and Augmented Reality. Now, as a PhD and in addition to what have been said before, he is also working on Machine Learning and Internet of Things.

Hamada EL KABTANE
Computer Science, Faculty of sciences SEMLALIA
Address: CADI AYYAD University, 40 000, MARRAKECH, MOROCCO
Phone: +212 604 239 424,
E-mail: elkabtanehamada@gmail.com

REFERENCES

Alonso-betanzos, A. (2007). Filter methods for feature selection. A comparative study. In *Intelligent Data Engineering and Automated Learning - IDEAL 2007* (pp. 178–187). https://doi.org/10.1007/978-3-642-04394-9

Alves, A. (2017). Stacking machine learning classifiers to identify Higgs bosons at the LHC. *Journal of Instrumentation*, *12*(5). https://doi.org/10.1088/1748-0221/12/05/T05005

Armbrust, M., Ghodsi, A., Zaharia, M., Xin, R. S., Lian, C., Huai, Y., … Franklin, M. J. (2015). Spark SQL: Relational Data Processing in Spark. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*, 1383–1394. https://doi.org/10.1145/2723372.2742797

Bahassine, S., Madani, A., Al-sarem, M., & Kissi, M. (2018). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2018.05.010

Burgos, C., Campanario, M. L., Pe??a, D. de la, Lara, J. A., Lizcano, D., & Mart??nez, M. A. (2017). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering*, *0*, 1–16. https://doi.org/10.1016/j.compeleceng.2017.03.005

Chaplot, D. S., Rhim, E., & Kim, J. (2015). Predicting student attrition in MOOCs using sentiment analysis and neural networks. *CEUR Workshop Proceedings*, *1432*, 7–12.

Chen, Z., Brandon, A., Gayle, C., Nicholas, E., Daphne, K., & J.Ezekiel, E. (2015). Who's benefiting from MOOCs, and why. *Harvard Business Review*, *25*.

Choudhury, S., & Bhowal, A. (2015). Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection. *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2015 - Proceedings*, (May), 89–95. https://doi.org/10.1109/ICSTM.2015.7225395

Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*, 6–14. https://doi.org/10.1145/2883851.2883931

Erel, I., Stern, L. H., Tan, C., & Weisbach, M. S. (2018). Selecting Directors Using Machine Learning. *Ssrn*. https://doi.org/10.2139/ssrn.3144080

Feng, W., Tang, J., & Liu, T. X. (2019). Understanding Dropouts in MOOCs.

Gao, C., Cheng, Q., He, P., Susilo, W., & Li, J. (2018). Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack. *Information Sciences*, *444*, 72–88. https://doi.org/10.1016/j.ins.2018.02.058

Goel, S., Sabitha, A. S., & Choudhury, T. (2019). *Analytical Analysis of Learners' Dropout Rate with Data Mining Techniques* (Vol. 841). Springer Singapore. https://doi.org/10.1007/978-981-13-2285-3

Gupta, R., & Sambyal, N. (2013). An understanding Approach towards MOOCs. *International Journal of Emerging Technology and Advanced Engineering*, *3*(6), 312--315. Retrieved from http://www.ijetae.com/files/Volume3Issue6/IJETAE_0613_52.pdf

Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout Prediction in MOOCs using Learner Activity Features. *ELearning Papers*, *37*(March), 1–10. Retrieved from https://oerknowledgecloud.org/sites/oerknowledgecloud.org/files/In_depth_37_1 (1).pdf

Jain, I., Jain, V. K., & Jain, R. (2018). Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification. *Applied Soft Computing Journal*, *62*, 203–215. https://doi.org/10.1016/j.asoc.2017.09.038

Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. *38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings*, 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458

Kabir, A., Ruiz, C., & Alvarez, S. A. (2014). Regression, Classification and Ensemble Machine Learning Approaches to Forecasting Clinical Outcomes in Ischemic Stroke. In *Biomedical Engineering Systems and Technologies* (Vol. 452, pp. 376–402). Springer International Publishing. https://doi.org/10.1007/978-3-662-44485-6

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Feature Subset Selection Problem using Wrapper Approach in Supervised Learning. *International Journal of Computer Applications*, *1*(7), 13–17. https://doi.org/10.5120/169-295

Khourdifi, Y., & Bahaj, M. (2018). Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms. In *International Symposium on Advanced Electrical and Communication Technologies (ISAECT)* (pp. 1–6).

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature Selection: A Data Perspective. *ACM Computing Surveys*, *50*. https://doi.org/10.1201/9781351070348

Liu, T., & Li, X. (2017). Finding out Reasons for Low Completion in MOOC Environment : An Explicable Approach Using Hybrid Data Mining Methods, (Meit), 376–384.

Liyanagunawardena, T. R., Parslow, P., & Williams, S. A. (2014). Dropout: MOOC Participants' Perspective. *Proceedings of the European MOOC Stakeholder Summit 2014*, 95–100. Retrieved from http://centaur.reading.ac.uk/36002/

Ly, A., Marsman, M., & Wagenmakers, E. (2018). Analytic posteriors for Pearson ' s correlation coefficient. *Statistica Neerlandica*, *72*(1), 4–13. https://doi.org/10.1111/stan.12111

Martinez-Espana, R., Bueno-Crespo, A., Timón, I., Soto, J., Munoz, A., & Cecilia, J. M. (2018). Air-pollution prediction in smart cities through machine learning methods: A case of study in Murcia, Spain. *Journal of Universal Computer Science*, *24*(3), 261–276.

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., … Talwalkar, A. (2015). MLlib: Machine Learning in Apache Spark. *Journal of Machine Learning Research*, *17*, 1–7. https://doi.org/10.1145/2882903.2912565

Mu, Y., Liu, X., & Wang, L. (2017). A Pearson's correlation coefficient based decision tree and its parallel implementation. *Information Sciences*. https://doi.org/10.1016/j.ins.2017.12.059

Naghibi, S. A., Ahmadi, K., & Daneshi, A. (2017). Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping. *Water Resources Management*, *31*(9), 2761–2775. https://doi.org/10.1007/s11269-017-1660-3

Nagi, S., & Bhattacharyya, D. K. (2013). Classification of microarray cancer data using ensemble approach. *Network Modeling and Analysis in Health Informatics and Bioinformatics*, *2*(3), 159–173. https://doi.org/10.1007/s13721-013-0034-x

Onah, D. F. ., Sinclair, J., & Boyatt. (2014). DROPOUT RATES OF MASSIVE OPEN ONLINE COURSES : BEHAVIOURAL PATTERNS MOOC Dropout and Completion : Existing Evaluations. *Proceedings of the 6th International Conference on Education and New Learning Technologies (EDULEARN14)*, 1–10. https://doi.org/10.13140/RG.2.1.2402.0009

Os, H. J. A. Van, Ramos, L. A., Hilbert, A., & Leeuwen, M. Van. (2018). Predicting Outcome of Endovascular Treatment for Acute Ischemic Stroke : Potential Value of Machine Learning Algorithms. *Frontiers in Neurology*, *9*(September), 1–8. https://doi.org/10.3389/fneur.2018.00784

Salcedo-Sanz, S., Cornejo-Bueno, L., Prieto, L., Paredes, D., & Garcia-Herrera, R. (2018). Feature selection in machine learning prediction systems for renewable energy applications. *Renewable and Sustainable Energy Reviews*. https://doi.org/10.1016/j.rser.2018.04.008

Sanchez-Gordon, S., & Luján-Mora, S. (2016). How could MOOCs become accessible? The case of edX and the future of inclusive online learning. *Journal of Universal Computer Science*, *22*(1), 55–81.

Sikora, R., & Al-Laymoun, O. (2014). A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms, *23*(1), 43–53. https://doi.org/10.4018/978-1-4666-7272-7.ch004

Talavera, L. (2005). An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering, 440–451. https://doi.org/10.1007/11552253_40

Urbanowicz, R. J., Meeker, M., Lacava, W., Olson, R. S., & Jason, H. (2018). Relief-Based Feature Selection : Introduction and Review. *Journal of Biomedical Informatics*, *85*, 189–203.

Vora, M. N. (2011). Hadoop-HBase for Large-Scale Data. In *International Conference on Computer Science and Network Technology* (pp. 601–605).

White, T. (2012). *Hadoop: The definitive guide*. (M. Loukides & M. Blanchette, Eds.), *Online* (3rd Editio). USA: O'Reilly Media, Inc. https://doi.org/citeulike-article-id:4882841

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior*, *58*(May), 119–129. https://doi.org/10.1016/j.chb.2015.12.007

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2017). Erratum: Corrigendum to "Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization" (Computers in Human Behavior (2016) 58 (119–129)(S074756321530279X)(10.1016/j.chb.2015.12.007)). *Computers in Human Behavior*, *66*, 409. https://doi.org/10.1016/j.chb.2016.08.051

Zhu, Y., Xie, C., Wang, G. J., & Yan, X. G. (2017). Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing and Applications*, *28*(s1), 41–50. https://doi.org/10.1007/s00521-016-2304-x

Zitlau, R., Hoyle, B., Paech, K., Weller, J., Rau, M. M., & Seitz, S. (2016). Stacking for machine learning redshifts applied to SDSS galaxies. *Monthly Notices of the Royal Astronomical Society*, *460*(3), 3152–3162. https://doi.org/10.1093/mnras/stw1454