A proposal for predicting and intervening on MOOC learners' performance in real time

Iván Pascual [1], Ruth Cobos [1]

[1] *Computer Science Department. Universidad Autónoma de Madrid, Spain*

Abstract

There is a lot of data from MOOCs, but their instructors cannot process that much information. While many learners end up dropping out of the course in which they enrolled in, a substantial problem in this context, their engagement data reveals their lack of interest even before they drop out. In order to make use of this information, we propose a Machine Learning approach to predict in real-time whether a learner would drop out or pass the MOOC, and a web-based dashboard approach to support this information and provide interventions over these learners. Using it in an asynchronous MOOC for 4 months, we predicted, with 0.93 F1-Score, the dropouts and passes from that period.

Keywords

Learning Analytics, Machine Learning, Massive Open Online Course, Prediction, Prescription, Intervention, Dashboard, Engagement.

1. Introduction and motivation

Recently, online courses have been increasing its popularity over the years; the courses known as "Massive Open Online Courses" (MOOCs) being one category of them, marking its relevance in the e-Learning discipline [1]. The Universidad Autónoma de Madrid (UAM) offers MOOCs since its entry in the edX Consortium (https://edx.org) in 2014, a platform for all universities to publish their course, and for all learners over the world to enrol in them, if they wish so. This led to the foundation of UAMx (https://uamx.uam.es), UAM's site for MOOCs. The approach presented in this article was tested in one of its courses: "Introduction to Development of Web Applications", also referred to as "WebApp MOOC" (https://uamx.uam.es/courses/course-v1:UAMx+WebApp+1T2019a/about); this also being the case of previous works that this one relies on ([2], [3], [4]) and which one of the authors of this article also worked in.

A recent field of study, the Learning Analytics (LA), proposes "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" [5], using software development, data mining, artificial intelligence, big data, and such, to this purpose.

As stated by several research studies, one of the main problems that affect MOOCs is the lack of interaction between learners and instructors, this leading to a decrease in learners' interest, finally resulting in the dropping out of the course by learners [6]. One of LA scopes is to lower this dropout rate, and keep the learners engaged with the course, helping them in the task of completing it.

1.1. The proposal

In this paper, we propose an approach to keep learners' engagement in MOOCs, and to prevent a potential dropout for a learner that was losing interest. In particular, we have a proposal that can be

[1]earning Analytics Summer Institute Spain (LASI Spain) 2022, June 20–21, 2022, Salamanca, Spain

EMAIL: ivan.pascuals@estudiante.uam.es (A. 1); ruth.cobos@uam.es (A. 2);

© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3238/paper4.pdf

broken down into two different approaches, forming together a single Prediction-Prescription system, which is as follows:

1. Analyze data from learners' activity in the course and determine if this learner's engagement is decreasing (Prediction).
2. Warn the instructor about the predictions, open ways to intervene on the cases, and automatically inform learners of their situation and suggest them how to keep up with the course (Prescription).

The structure of this article is as follows: we now review, in the next section, the state of the art on dropout prediction in MOOCs, and the system that serves as a background to our proposal (edX-LIMS, [2], [3]). In section 3, the Machine Learning on real-time predictions approach is presented. Then, in section 4, learners' and instructors dashboards are described. Finally, the article ends briefly presenting current results (section 5) and the conclusions (section 6).

2. State of the art and background

2.1. Learning Analytics

Learning Analytics (LA), as stated, is a field of study that encourages learning data recollection and analysis to help learners and their academical environment [7]. The Society of Learning Analytics Research (SoLAR), organizes events as the Learning Analytics Summer Institute (LASI, starting on 2013); and the Spanish Network of Learning Analytics (SNOLA) holds the LASI for Spain, which has been promoting works in this area and congregating professionals of both computer science and education under this one, single goal: to use learners' data analysis in benefit of them.

Related research lines, presented as achievements from the SNOLA network [8] and the GHIA group at UAM [9], include (but are not limited to):

- Predictive analysis for dropout and students at risk ([4], [6], [10], [11]) using Ma-chine Learning (ML) algorithms, to improve learner retention and engagement, as predictive systems.
- Visual analysis ([12], [13]) using dashboard methodologies to describe the data that is collected and drive the decisions taken by both instructors and learners.
- Prescriptive analysis ([15], [16]) as the methodologies to provide personalized feed-back to learners in educational environments.

As with the predictive analysis topic, we found more theoretical research studies (for example, [17], [18]) than real systems that apply ML techniques. As such, our proposal would be one of these, that feeds on these studies to bring a system focused on the end-user (teachers and learners).

2.2. edX Learning and Intervention Monitoring System: the initial context

edX-LIMS (which was developed from edX-LIS, [17]) is a "Web-based Learning Analytics System, which provides an intervention strategy on the learners' learning and the monitoring of the mentioned strategy by the instructors" [3]. It is currently deployed on the WebApp MOOC, although its design is intended to work for any edX MOOC. It is the base system for instructors to collect and measure data from the learners' activity, achieved by log processing and a database schema. Currently, edX only allows to do evaluations (assessments) to learners enrolled in the verified itinerary, those being the ones who pay a fee and can obtain an official certificate at the end of the course. edX-LIMS's code is written in Python and the data is stored in a non-relational, MongoDB database using different collections. The web interface is generated with the Dash framework (https://plotly.com/dash/), based on Flask, with which the final server is based.

The system has only three possible users: learners (who can view its own data), instructors (who can view all learners' and overall course data), and an admin (who maintains the application). The most important service of edX-LIMS for this article is the Data Processing Service. It collects the data following three phases: i) firstly, the log files provided by edX are processed; ii) then, with this processing, the indicators of the course (such as the time spent watching videos or answering assessments) are calculated and iii) finally they are stored in the MongoDB database. The rest of them include visualization of dashboards with statistics about the course (to both the instructors and learners), a mailing tool to quickly send e-mails to some specific learners and an engagement monitoring service.

3. The real-time predicting on learners' performance proposal

We propose a methodology to predict, given the accumulative indicators of activity of a learner and their grades in a MOOC in real-time, whether that learner is going to certificate (pass the course) or to drop out, using an ML approach. To that end, this section describes how the target was defined, our specific dataset available and the model selection and exploitation process.

3.1. Defining the target

The first problem we found on predicting this target was to define it. WebApp is an asynchronous MOOC, meaning that a learner can be inactive for any number of days, and then, come back. Thus, formally speaking, no learner could ever be considered a "drop out", so we may consider a learner to have dropped out when he or she surpasses a given number of consecutive inactivity days and the problem at hand is to define this number.

Some approaches on this matter ([4], [10]) use percentiles to measure the overall tendency, and we decided to follow this line. We may compute percentiles on learners that ended up passing the course, and get the results shown in Table 1. To not overestimate this number, but to capture at least the tendency of a majority of passed learners, we shall define 98 (90-percentile) as the number of consecutive inactivity days after which we consider a learner to have dropped out.

Table 1
Percentiles on most consecutive inactivity days for learners who ended up passing the course.

Percentile (%)	10	25	50	70	80	90	95	100
Most consecutive inactivity days	0	2.5	8	21	35.3	98	213	695

Moreover, we only consider a learner to have dropped out if that learner has not passed yet, so the two conditions for being labelled as a dropout are as follows:

1. The learner has been inactive for 98 or more days, consecutively.
2. The learner has a grade lower than 0.5, as the WebApp constrains (i.e., has not passed yet).

The opposite of "drop out" is widely considered as "not drop out" (i.e., as a binary variable), as shown in Figure 1. If the learner passes or not, is studied independently, as another (binary) target variable. Thus, we may be tempted to predict down two tar-gets: dropout and pass, as two separated classification problems.

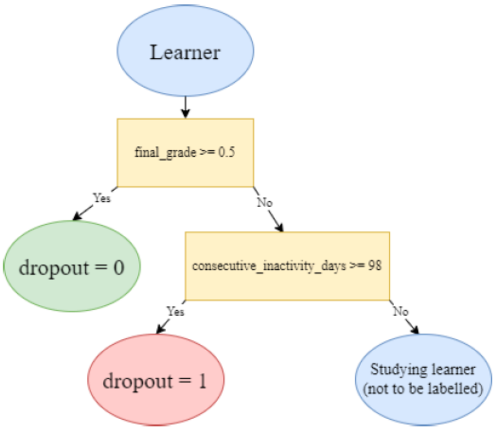


Figure 1: Decision Tree on conditions of a "Dropout" target (modified from [4]).

However, we shall briefly see that, in our prediction context, a negative (0) dropout must mean a positive (1) pass, and that a negative pass must mean that either the learner dropped out (positive on dropout) or that is still studying the course. Notice from Figure 1 that, if a learner does not obtain a positive dropout, it means that the learner: i) is studying the course, or ii) has passed the course. The learner from the first case was not to be labelled as a dropout at all, but the second one would obviously be labelled as a positive in pass.

Let us elaborate further on the "pass" and "studying" labels; we label a learner as a "pass" simply if:

1. The learner has a grade of 0.5 greater.

And, finally, to be studying the course means the negation of the two before, in other words: to not have passed, but also to not have dropped out. In essence:

1. The learner has been inactive for less than 98 days, consecutively.
2. The learner has a grade lower than 0.5.

Table 2 shows the truth tables of two theoretical "Pass" and "Dropout" classes that follow all these conditions, and the interpretation from their combination. As the "stud-ying" label is not to be predicted, these interpretations leave us with a single, categorical variable. The final target variable is shown by Table 3, along with their conditions.

Table 2
Truth table and interpretations of a combination "Pass" and "Dropout" classes. "N" means "Negative" (0), and "P" means "Positive" (1). In bold, the only labels that make sense to predict in this context.

"Pass"	"Dropout"	Interpretation	Label
N	N	A studying learner, as they have not yet passed, nor dropped out.	Studying
N	**P**	**A learner that has dropped out.**	**Dropout**
P	**N**	**A learner that has passed the course.**	**Pass**
P	P	An impossible combination by the conditions stated in this section.	None

Table 3
Truth table and interpretations of a combination "Pass" and "Dropout" classes. "N" means "Negative" (0), and "P" means "Positive" (1). In bold, the only labels that make sense to predict in this context.

Dropout (0)	Pass (1)
final_grade < 0.5 **consecutive_inactivity_days >= 98**	final_grade >= 0.5

3.2. The learners' data

Moving forward, we now explain the structure of the data and how it is preprocessed to train a ML algorithm. The data used in the predictions come directly from the logs supplied by edX. Then, Course Data Processing Service of the edX-LIMS process these logs into the Mongo database, wherefrom it is then read. The WebApp MOOC started on April 9, 2019, as well as our data.

Our specific dataset comes from two different collections of this database, namely, the collection where accumulative indicators (AI) for each learner and day are stored, and the collection where grades for each evaluation (GEA) for each learner and week are stored. After testing on different features (with the methodology explained later in Section 3.3), on the final model (described in Section 3.4), we got the results on feature importance that Figure 2 shows.

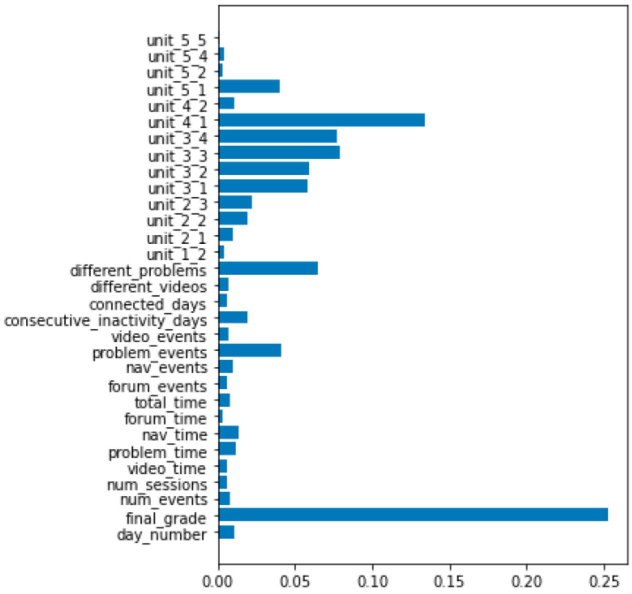


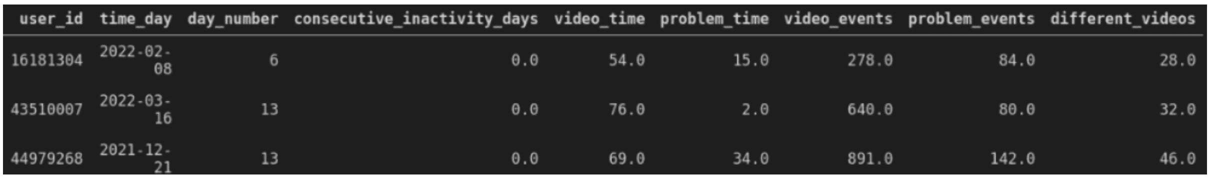
Figure 2: Feature importance on all GEA ("unit_*" variables) and AI (rest) from data. Detail on AI's can be found at [19], but is summarized below.

Obviously, the grade that the learner holds (final_grade) is the most important feature. We observed that the forum variables and the number of days connected were the lowest, and so we discarded them.

As a result, each collection in the prediction process has the following variables:

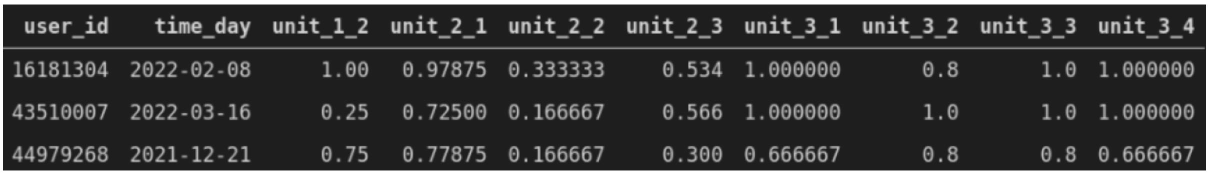
- AI: Total grade, number of days since sign up, number of consecutive inactivity days, and the number of both page loads and time spent for: interactive sessions, videos, evaluations, and overall navigation.
- GEA: The grade for each of the evaluations of the course.

Data between these collections are then merged to make up our dataset. Note that because AI is daily, and GEA weekly, the resulting dataset turns out logically weekly. This is no problem, because, as the AI indicators are accumulative, data between weeks are added up to the next. We may call a row of our dataset a set of all AI and GEA, for a learner and a weekly date. In other words: for each enrolled learner and week, we have their AI and GEA at that time. We show the structure of such merged rows in both Figure 3 and 4. In Figure 3, only 7 of the total AI are shown. In Figure 4, only 8 of the total GEA are shown. Note that the Figure 4 is the continuation of Figure 3; i.e., each row of the dataset has both the AI and GEA, for each learner and week.



user_id	time_day	day_number	consecutive_inactivity_days	video_time	problem_time	video_events	problem_events	different_videos
16181304	2022-02-08	6	0.0	54.0	15.0	278.0	84.0	28.0
43510007	2022-03-16	13	0.0	76.0	2.0	640.0	80.0	32.0
44979268	2021-12-21	13	0.0	69.0	34.0	891.0	142.0	46.0

Figure 3: Example rows from our resulting dataset, showing only 7 of the AI for 3 different users on 3 different days.



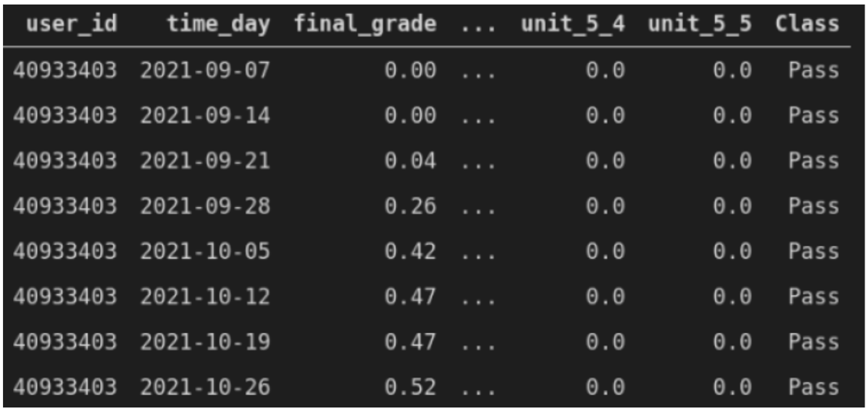
user_id	time_day	unit_1_2	unit_2_1	unit_2_2	unit_2_3	unit_3_1	unit_3_2	unit_3_3	unit_3_4
16181304	2022-02-08	1.00	0.97875	0.333333	0.534	1.000000	0.8	1.0	1.000000
43510007	2022-03-16	0.25	0.72500	0.166667	0.566	1.000000	1.0	1.0	1.000000
44979268	2021-12-21	0.75	0.77875	0.166667	0.300	0.666667	0.8	0.8	0.666667

Figure 4: Continuation of the example rows of our dataset, showing the grade of each learner, for each day, on only 8 of the evaluations. These 3 users are the same from Figure 3.

Appended to the end of these rows is the class or label assigned to each learner, but not to each day, inferred from the data in all the learners' rows by the conditions stated in section 3.1. In other words: if, for a given learner, there exists such a row that their final_grade is greater or equal to 0.5, then a "Pass" is labelled and propagated for all their previous rows. If, however, the "Dropout" conditions are met, then all their rows are labelled as such. An example of this is shown on Figure 5.

With this, we label all the activity of a passed (or a dropout) learner as such, thus making the algorithm learn that a learner that ends up passing (or dropping out) could have the indicators, the grade, the days enrolled, etc., that the row shows. Using this approach, we have the number of Dropout and Pass rows shown in Table 4.

If neither of the "Pass" or "Dropout" labels are satisfied, then none is assigned, and this row of data would be the subject of our predictions: as this is a studying learner for that day, we will want to predict if it would become a "Pass" or a "Dropout", based on the state of affairs – the learners' data for that date, its "row".



```
user_id    time_day  final_grade  ...  unit_5_4  unit_5_5  Class
40933403  2021-09-07         0.00  ...       0.0       0.0   Pass
40933403  2021-09-14         0.00  ...       0.0       0.0   Pass
40933403  2021-09-21         0.04  ...       0.0       0.0   Pass
40933403  2021-09-28         0.26  ...       0.0       0.0   Pass
40933403  2021-10-05         0.42  ...       0.0       0.0   Pass
40933403  2021-10-12         0.47  ...       0.0       0.0   Pass
40933403  2021-10-19         0.47  ...       0.0       0.0   Pass
40933403  2021-10-26         0.52  ...       0.0       0.0   Pass
```

Figure 5: An example on the "label propagation". Learner with ID 40933403 passes the course on October 26, 2021. As there exists such a row that its grade is greater than 0.5, then all their previous records are labelled as a "Pass".

Table 4
Number of rows of the training dataset labelled with Dropout and Pass labels through the label propagation.

Dropout	Pass
49927	40642

3.3. Model selection and tests

Now the train-test process is explained, and the results of the model selection are shown. We split our dataset in a temporal way, as in Table 5.

Table 5
Periods of time used in training and testing the models.

Training dataset	Test dataset
2 years and a half of data	2 months of data

The reason for such a difference between the size of the training and test dataset is for the exploitation process. When we predict on actual learners, we will train the model with all the data from the course, but the last week (as explained in Section 3.4). One week for testing is too small to get generalized results, but classic partitions as 80-20 or 70-30 percent of data would not test the exploitation process fairly. We decided, then, to test on 2 months of data: 8 weeks at most.

We then fitted Decision Tree, SVC, Random Forest, Gradient Boosting and Ada-Boost algorithms to the training dataset, and tested against the test dataset. We measured the precision, recall and F1-Score of them all, and the results are shown in Table 6. Given such results, we chose the Random Forest algorithm for their overall good performance compared to the others, and fast execution thanks to its capacity to parallel process. Finally, we performed a grid search on hyperparameters, and found a better score to a rather simple set: 100 decision tree estimators, 1 minimum samples leaf, 2 minimum samples split and no constrains on maximum tree depth.

3.4. Final model and prediction process

After this model selection process, the next step of this proposal is how to exploit an ML prediction algorithm in this context. Weekly, WebApp MOOC logs are processed by edX-LIMS and data is extracted; in particular, accumulative indicators on learners' engagement with the course and their grades.

Table 6
Results of testing different algorithms on the training dataset.

Averages	Decision Tree	SVC	**Random Forest**	Gradient Boosting	AdaBoost
Precision	0.81	0.6	0.89	0.84	**0.91**
Recall	0.81	0.58	**0.89**	0.82	0.78
F1-Score	0.8	0.53	**0.88**	0.80	0.74

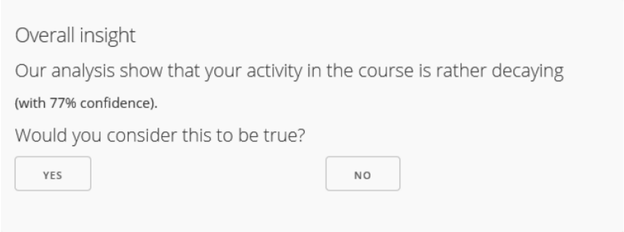
Our approach proposes to train a Random Forest model upon all data that can be labelled and predict on all that cannot. If a learners' row can be labelled, it means that either that learner has passed or dropped out. If it cannot be labelled, however, it means that this learner is still studying the course and we may predict on real-time if he or she will pass or not. If a learner has a majority of "Pass" predictions over their studying, we'll say that he or she will pass (and the same goes for a majority of "Dropout" pre-dictions). These are our target learners, and the workflow, for each week, with this approach, is as follows:

1. A week's engagement and grade data ("the week", in this list) are processed through edX-LIMS. Each learner has a unique row of data for the week.
2. The model is (re)trained upon all data that can be labelled (i.e., with learners' data that already passed or dropped out), come it from the week's data or not.
3. The model is tested upon all learners' data, from the week, that cannot be labelled (i.e., learners' that are still studying). The model predicts "Pass" or "Dropout" for each learner, with their week's accumulative data, along with a probability.
4. The model's predictions, probabilities, and Shapley explanations on the predictions are stored into the Mongo database for later presentation.

4. The web-based intervening system proposal

4.1. The learners' view

Our first and most important focus on the matter of intervention is to inform learners of the analysis. We do so sending them an e-mail with an unique link to a web-based dashboard where they can access the mentioned analysis. We may call the ML predic-tions a "statistical analysis" (or, simply put, "analysis"), especially when talking to learners, because the "prediction" term can result a greedy one from their perspective (“we predict that you will…”). We also avoid using “dropout”, as it can be demoralizing. This is crucial to this first intervention to be effective: we want our learners to get along with our analysis, not to reject them. To this end, we may present the result of our predictions as Figure 6 shows, with a feedback question that instructors may take in account. A history of these analysis is also graphed, as Figure 7 shows, so the learners can also view past analysis.



Overall insight

Our analysis show that your activity in the course is rather decaying

(with 77% confidence).

Would you consider this to be true?

Figure 6: An example of a “Dropout” prediction message to the learner, though the word “dropout” is not used.

In the prediction process, we also extract the Shapley values [20] of the predictions to estimate how much each accumulative indicator contributed to which class (“Drop-out” or “Pass”). Another form of intervention is the following: we intend to guide the learners progress showing them these values in an understandable way, such as they may see in which of them may be flawing and doing good. We do not store Shapley values of the GEA variables, as the learners cannot change them after they obtained them. What they can change, however, are the AI, by changing their behavior. An example of how this information is graphed is shown in Figure 8.

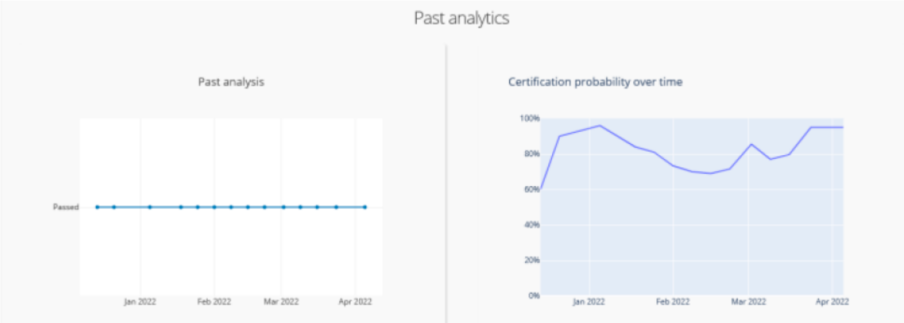
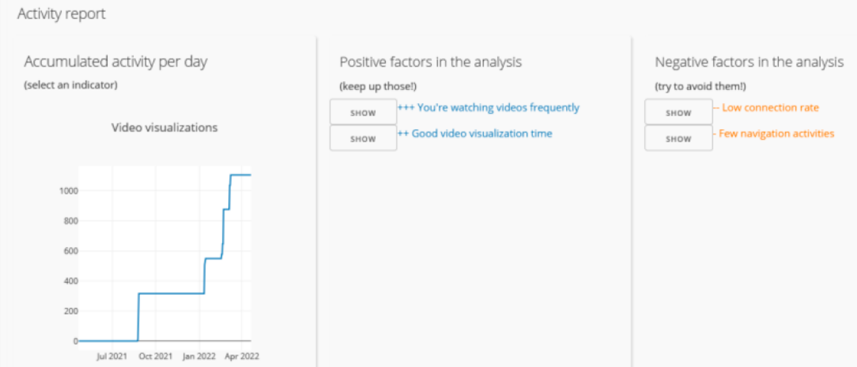


Figure 7: A graph of past analysis done to a learner, from the learner’s view.



Activity report

Accumulated activity per day

(select an indicator)

Positive factors in the analysis

(keep up those!)

+++ You're watching videos frequently

++ Good video visualization time

Negative factors in the analysis

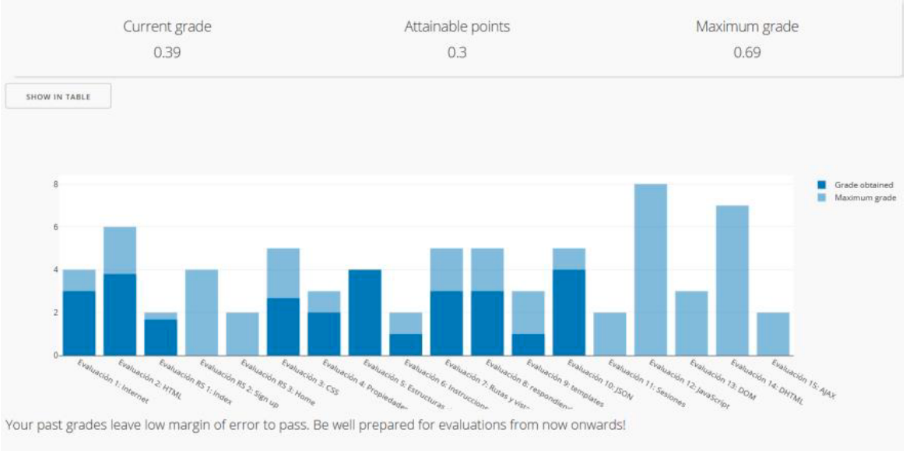
(try to avoid them!)

-- Low connection rate

- Few navigation activities

Figure 8: “Activity report”, as the showing of Shapley values on the last week’s prediction is called for the learner. More “+” or “-“ mean more significance towards each class.

In addition to all this ML techniques and predictions, we also integrate to our proposal a simple grade monitor that shows the current grade of the learner, their still attainable points of evaluations not yet done, and the resulting maximum grade attainable (the sum of the two before). As a learner progresses through the course, they may encounter that they mathematically cannot pass the course anymore because they had performed too poorly on past evaluations. To avoid this as much as possible is the objective of this tool.



Current grade	Attainable points	Maximum grade
0.39	0.3	0.69

Your past grades leave low margin of error to pass. Be well prepared for evaluations from now onwards!

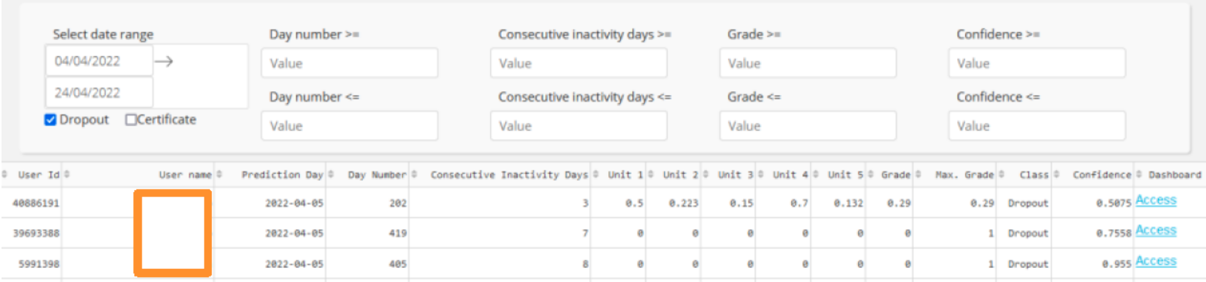
Figure 9: Grades obtained (dark blue) vs. maximum grades (light blue) for the learners. This learner has only 0.3 points left to add up to a maximum of 0.69.

With the information shown in Figure 9 for the learners, they can be conscious of their situation, and prepare better for the evaluations to come if they feel that they may be on the edge of passing the course (as it is warned on that Figure).

With all this data, learners can have a better view of how they are progressing in the course and how they can improve, while also giving them a sense of control over their performance; and that, itself, is an intervention through information.

4.2. The instructors' view

Even so, instructors, from their part, can be informed of the predictions and progres-sion of the learners through their own web-based dashboard. The first table shown, as in Figure 10, is the "Last week's summary", in which they can watch the last week's predictions on studying learners, the confidence on the prediction, their grades, consec-utive inactivity days, days since enrollment and a link to that learners' dashboard (to examine their point of view, if necessary). They may also filter past predictions or some of the mentioned values.



Select date range
04/04/2022 →
24/04/2022
☑ Dropout □Certificate

Day number >= Value
Day number <= Value
Consecutive inactivity days >= Value
Consecutive inactivity days <= Value
Grade >= Value
Grade <= Value
Confidence >= Value
Confidence <= Value

User Id	User name	Prediction Day	Day Number	Consecutive Inactivity Days	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Grade	Max. Grade	Class	Confidence	Dashboard
40886191	[redacted]	2022-04-05	202	3	0.5	0.223	0.15	0.7	0.132	0.29	0.29	Dropout	0.5075	Access
39693388	[redacted]	2022-04-05	419	7	0	0	0	0	0	0	1	Dropout	0.7558	Access
5991398	[redacted]	2022-04-05	405	8	0	0	0	0	0	0	1	Dropout	0.955	Access

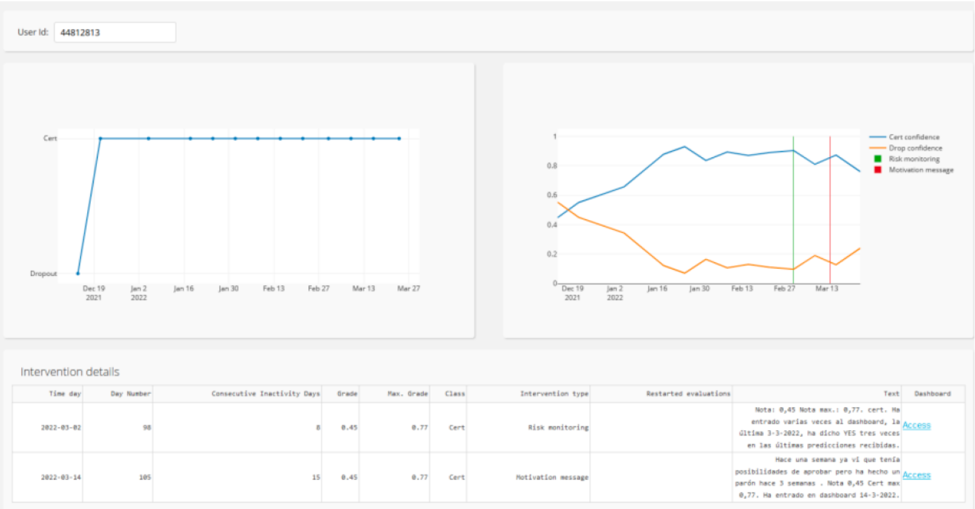
Figure 10: Predictions of last week section for instructors. Usernames hidden.

The next graphs (shown in Figure 11) let instructors to watch the past predictions for a specific learner (identified by their user ID) and all their registered interventions on that learner. The intervention mark is complemented with some indicators from the day the intervention was done (as the consecutive inactivity days the learner had when the intervention was made, or their grade), which is a useful way to recall the situation of a certain learner and even quantify if the intervention was successful or not.

Instructors also have access to the information of grades and maximum grades, for all learners, so they may take them into account, and, specially, be warned of the ones who ended up failing to personally evaluate their situation. Such information is summarized in a table like Figure 12 shows, in which the individual grade for each evaluation, the total grade, points still attainable and maximum grade can be viewed. Instructors can filter for learners that failed, so they can quickly intervene if they wish so.

Finally, to annotate all these interventions made on the learners and keep track of their progress, a simple annotator on interventions is also implemented for the instructors. The interface is shown in Figure 13. Experience on the WebApp MOOC has taught us at least four types of interventions:

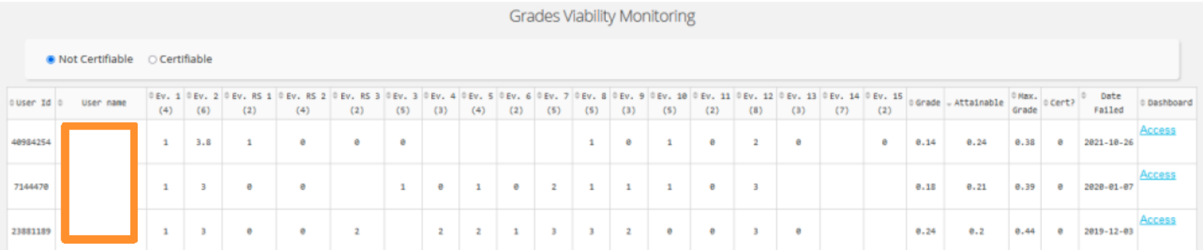
1. Sending motivational messages: in the case of a learner who is losing interest in the course. This is done by the instructor, by a personal e-mail, sent through the system.
2. Offering the possibility of reinitializing some evaluations: in the case that the learner cannot pass the course taking into account the actual grade and their still attainable points of evaluations not yet done. This takes the form of an e-mail sent by the in-structor, through the system, describing the situation and asking them if they want some evaluations to be reinitiated.
3. Executing the reinitialization of the evaluations: in the case that the learner confirms that he or she would like to repeat some evaluations.
4. Monitoring for possible risk: in the case that the learner could need one of the pre-vious mentioned interventions in the future.



Intervention details

Time day	Day Number	Consecutive Inactivity Days	Grade	Max. Grade	Class	Intervention type	Restarted evaluations	Text	Dashboard
2022-03-02	98	8	0.45	0.77	Cert	Risk monitoring		Nota: 0,45 Nota max.: 0,77. cert. Ha entrado varias veces al dashboard, la última 3-3-2022, ha dicho YES tres veces en las últimas predicciones recibidas.	Access
2022-03-14	105	15	0.45	0.77	Cert	Motivation message		Hace una semana ya vi que tenía posibilidades de aprobar pero ha hecho un parón hace 3 semanas . Nota 0,45 Cert max 0,77. Ha entrado en dashboard 14-3-2022.	Access

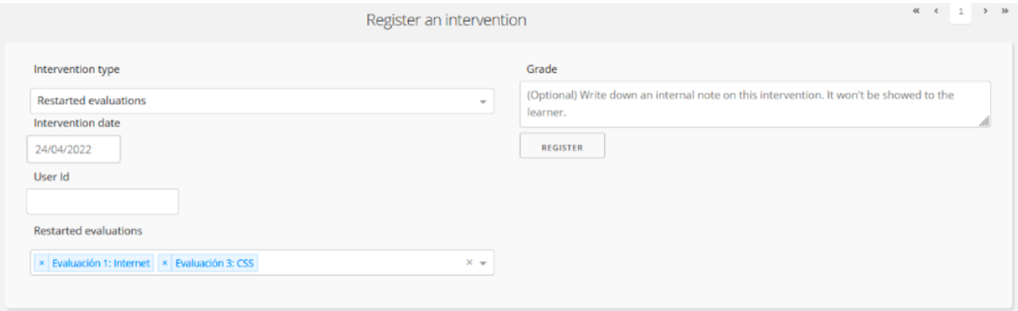
Figure 11: An example of a progress graph from an instructor’s perspective. It can be seen two interventions done on this learner, which, in account of, ended up passing a week later.


Grades Viability Monitoring

● Not Certifiable ○ Certifiable

User Id	User name	Ev. 1 (4)	Ev. 2 (6)	Ev. RS 1 (2)	Ev. RS 2 (4)	Ev. RS 3 (2)	Ev. 3 (5)	Ev. 4 (3)	Ev. 5 (4)	Ev. 6 (2)	Ev. 7 (5)	Ev. 8 (5)	Ev. 9 (3)	Ev. 10 (5)	Ev. 11 (2)	Ev. 12 (8)	Ev. 13 (3)	Ev. 14 (7)	Ev. 15 (2)	Grade	Attainable	Max. Grade	Cert?	Date Failed	Dashboard
40984254	[redacted]	1	3.8	1	0	0	0					1	0	1	0	2	0		0	0.14	0.24	0.38	0	2021-10-26	Access
7144470	[redacted]	1	3	0	0		1	0	1	0	2	1	1	1	0	3				0.18	0.21	0.39	0	2020-01-07	Access
23881189	[redacted]	1	3	0	0	2		2	2	1	3	3	2	0	0	3	0			0.24	0.2	0.44	0	2019-12-03	Access

Figure 12: Grade Monitoring table for instructors, filtered by learners that failed in the past. Blank means “Not attempted”. Usernames hidden.


Register an intervention

Intervention type: Restarted evaluations

Intervention date: 24/04/2022

User Id:

Restarted evaluations: × Evaluación 1: Internet × Evaluación 3: CSS

Grade: (Optional) Write down an internal note on this intervention. It won't be showed to the learner.

Figure 13: Intervention registration tool.

These four types are implemented as a drop-down menu on the interface shown in Figure 12, and then, each intervention is registered to review as in Figure 10.

5. Results and limitations

Our proposal on real-time predicting on learners' performance has been functionating since December 14, 2021, on WebApp MOOC. About four months later (April 28, 2022), 22 learners have passed this course, and 21 of them have been predicted to do so. It also has been the case that 32 learners dropped out the course, accumulating more than 98 days of consecutive inactivity whilst our model was predicting, and 30 of them were also predicted. These were mostly learners that were already enrolled in, before December 14. These numbers compute the practical metrics showed in Table 7 at the time of writing this article (April 28, 2022).

Table 7
The real precision, recall and F1-Score of the model, after about four months.

Precision	Recall	F1-Score
0.91	0.95	0.93

At the time of writing, up to 52 interventions were registered, done as a result of the predictions. Of these, we can attest to at least 3 of the 22 learners to have passed thanks to them. More than 500 accesses in different days were also registered, from different learners, were made to their dashboards, and we have yet to measure more on the impact of our proposal, as time passes.

As for the limitations, firstly, this ML approach only predicts on the data from one record (or one week), ignoring the past records of the learner. This leads to errors on the predictions, especially when few data are available for the learner for the first weeks. We have observed that this approach makes actual sense from the prediction when the data from the third week is processed; it is then when, clearly, a learner that has done nothing is predicted to drop out. Secondly, data from edX is sent weekly, which is a massive slowdown to this approach, which would benefit from daily data, as the interventions could be more interactive, and dropouts predicted earlier.

6. Conclusions

In conclusion, as a response to an alarmingly rate of dropout on MOOCs learners, usually motivated by a feeling of isolation on the completion of the course or motivation, we developed a Machine Learning system which uses learners' activity data as input to an Artificial Intelligence model which then statistically calculates whether or not any learner is likely to drop out of the course or going to pass the course (obtaining the certificate). Then, the system warns instructors and invite them to realize interventions to learners at risk of losing interest in the course and dropping. The proposed system supports these services to learners and instructors through the generation of web-based dashboards.

More in detail, the presented approach in this article contributes with:

1. Firstly, the analysis and predictions over learners' performance based on their engagement data, in real-time, through the workflow of training and computing pre-dictions with a supervised learning, Machine Learning algorithm, following a Learning Analytics approach.
2. Secondly, the intervention system through web-based dashboard. We implement a dashboard for instructors, through which the predictions and information about the course is shown, and where they can register their interventions. Learners also have their own dashboards to see their predictions and grades, so they can view how they are progressing.
3. Finally, the experiences of applying a real Learning Analytics system on a real course, with feedback from learners and engagement data with their new dashboards.

Our proposal is, in summary, a system that goes from predicting to prescribing, helping to intervene over learners that may lack motivation to complete their course and may be, otherwise, totally missed by the instructors. Our call is for such systems to be developed and integrated by a LA community that has the knowledge to materialize these solutions. Theoretical studies are necessary for the background of such systems, but we need to transport all these useful conclusions to the learners' environment, so they can benefit of all this knowledge, and make their learning a more interactive and satisfying one.

7. Acknowledgments

This work has been co-funded by the Madrid Regional Government through the e-Madrid-CM Project under Grant S2018/TCS-4307, a project which is co-funded by the European Structural Funds (FSE and FEDER). This research has been co-funded by the National Research Agency of the Spanish Ministry of Science, Innovation and Universities under project grant RED2018-102725-T (SNOLA). And, this research has been co-funded by the National Research Agency of the Spanish Ministry of Science and Innovation under project grant PID2019-105951RB-I00 (In-diGo!).

8. References

[1] By the Numbers: MOOCs During the Pandemic. Class Central. https://www.classcentral.com/report/mooc-stats-pandemic/, last accessed 2022/03/25.

[2] Soberón, J.: Sistema informático de apoyo a las analíticas para el aprendizaje (Learning Analytics) para entornos educativos on-line. Trabajo de Fin de Grado. Universidad Autónoma de Madrid. Madrid (2020).

[3] Cobos, R., Soberón, J.: A proposal for Monitoring the Intervention Strategy on the learning of MOOC learners. CEUR Conference Proceedings. LASI-Spain 2020. Learning Analytics Summer Institute Spain 2020: Learning Analytics. Time for Adoption?. (online) Valladolid, Spain. http://ceur-ws.org/Vol-2671/paper07.pdf

[4] Herrero, G.: Sistema informático para la predicción de certificado y abandono en entornos educativos en línea. Trabajo de Fin de Grado. Universidad Autónoma de Madrid. Madrid (2021).

[5] Lang, C., Siemens, G., Wise, A., Gasevic, D.: Handbook of Learning Analytics. Society for Learning Analytics Research (SoLAR), 2017. doi: 10.18608/hla17.

[6] Moreno-Marcos P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., Kloos C. D.: Prediction in MOOCs: A Review and Future Research Directions. IEEE Trans. Learn. Technol., vol. 12, no. 3, pp. 384–401, 2019, doi: 10.1109/TLT.2018.2856808.

[7] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," WIREs Data Min. Knowl. Discov., vol. 10, no. 3, May 2020, doi: 10.1002/widm.1355.

[8] Martínez Monés, A., et al.: Achievements and challenges in learning analytics in Spain: The view of SNOLA," Revista Iberoamericana de Educación a Distancia, vol. 23, no. 2, p. 187, (July 2020). doi: 10.5944/RIED.23.2.26541.

[9] Alamán, X., Carro R. M., Cobos, R., et al.: Modelado de Estudiantes, Analítica de Aprendizaje, Aten-ción a la Diversidad y e-Learning. IE Comunicaciones: Revista Iberoamericana de Información Educa-tiva, ISSN-e 1699-4574, no. 30 (Julio-Diciembre), 2019, 78–89.

[10] Moreno-Marcos, P. M.: Analítica del aprendizaje para la predicción en escenarios educativos heterogé-neos. Tesis Doctoral. Universidad Carlos III de Madrid, Madrid (2020).

[11] Alcolea, J. J., Ortigosa, A., Carro, R. M., & Blanco, O. J. (2020). Best Practices in Dropout Prediction: Experience-Based Recommendations for Institutional Implementation. In D. Glick, A. Cohen, & C. Chang (Ed.), Early Warning Systems and Targeted Interventions for Student Success in Online Courses (pp. 301-323). IGI Global. https://doi.org/10.4018/978-1-7998-5074-8.ch015

[12] Vázquez-Ingelmo, A., García-Peñalvo, F. J., Therón R.: Automatic generation of software interfaces for supporting decision-making processes. An application of domain engineering and machine learning. ACM International Conference Proceeding Series, pp. 1007–1011, (October 2019), doi: 10.1145/3362789.3362923.

[13] Verbert, K. et al., Learning dashboards: An overview and future research opportunities, Pers. Ubiquitous Comput., vol. 18, no. 6, pp. 1499–1514, Nov. 2014, doi: 10.1007/s00779-013-0751-2.

[14] Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., Mirriahi, N.: Using learning analytics to scale the provision of personalised feedback. Br. J. Educ. Technol., vol. 50, no. 1, pp. 128–138. (Jan. 2019), doi: 10.1111/bjet.12592.

[15] Iraj, H., Fudge, A., Faulkner, M., Pardo, A., Kovanović, V.: Understanding Students' Engagement with Personalised Feedback Messages. Proc. LAK 2020, pp. 438–447. doi: 10.1145/3375462.3375527.

[16] Topali, P., Ortega-Arranz, A., Er, E., Martínez-Monés, A., Villagrá-Sobrino, S. L., Dimitriadis, Y.: Exploring the problems experienced by learners in a MOOC implementing active learning pedagogies. 11475 LNCS. pp. 81–90 (May. 2019), doi: 10.1007/978-3-030-19875-6_10.

[17] Cobos, R., Ruiz-Garcia, J. C.: Improving learner engagement in MOOCs using a learning intervention system: A research study in engineering education. Computer Applications in Engineering Education. 29(4), 733–749 (2021). https://doi.org/10.1002/cae.22316.

[18] A. Larrabee Sønderlund, E. Hughes, and J. Smith, "The efficacy of learning analytics interventions in higher education: A systematic review," Br. J. Educ. Technol., vol. 50, no. 5, pp. 2594–2618, Sep. 2019, doi: 10.1111/bjet.12720.

[19] R. Cobos and L. Olmos, "A Learning Analytics Tool for Predictive Modeling of Dropout and Certificate Acquisition on MOOCs for Professional Learning," 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2018, pp. 1533-1537, doi: 10.1109/IEEM.2018.8607541.

[20] Shapley, L. S.: A Value for n-Person Games. Contributions to the Theory of Games (AM-28), Volume II, pp. 307–318 (May 2016). doi: 10.1515/9781400881970-018.