Early Predictor for Student Success Based on Behavioural and Demographical Indicators

Efthyvoulos Drousiotis[1], Lei Shi[2], Simon Maskell[1]

[1] Department of Electrical Engineering & Electronics, University of Liverpool, Liverpool, UK
{e.drousiotis, s.maskell}@liverpool.ac.uk
[2] Department of Computer Science, Durham University, Durham, UK
lei.shi@durham.ac.uk

Abstract. As the largest distance learning university in the UK, the Open University has more than 250,000 students enrolled, making it also the largest academic institute in the UK. However, many students end up failing or withdrawing from online courses, which makes it extremely crucial to identify those "at risk" students and inject necessary interventions to prevent them from dropping out. This study thus aims at exploring an efficient predictive model, using both behavioural and demographical data extracted from the anonymised Open University Learning Analytics Dataset (OULAD). The predictive model was implemented through machine learning methods that included BART. The analytics indicates that the proposed model could predict the final result of the course at a finer granularity, i.e., classifying the students into Withdrawn, Fail, Pass, and Distinction, rather than only Completers and Non-completers (two categories) as proposed in existing studies. Our model's prediction accuracy was at 80% or above for predicting which students would withdraw, fail and get a distinction. This information could be used to provide more accurate personalised interventions. Importantly, unlike existing similar studies, our model predicts the final result at the very beginning of a course, i.e., using the first assignment mark, among others, which could help reduce the dropout rate before it was too late.

Keywords: MOOCs · Virtual Learning Environment · Learning Analytics · Behavioural Analytics · Machine Learning · Prediction, BART

1 Introduction

Online learning offers a convenient alternative for everyone to learn on-demand. According to Class Central Report [1], more than 180 million students have enrolled in online learning courses, in particular, MOOCs (Massive Open Online Courses). Yet, one of the well-known challenges in online learning, especially in the context of MOOCs, is student retention. Studies, e.g. [2], show normally only 5% ~ 15% of the students who have registered for a MOOC finally complete it. Luckily, the massive data tracked on online learning platforms, so-called *Educational Big Data*, offers great opportunities to explore how students learn online thus providing insight into

(dis)engagement patterns. In fact, many studies have been conducted to predict student dropout, using techniques through statistical modelling [3] to machine learning [4,5].

However, most studies, e.g. [4], [6-8], proposed their predictive models using the learning activity data of a whole course, which are not particularly useful in terms of helping the *current* students, as the predictions are only made after the course has completed. A few studies did aim at an earlier prediction using the very first/early data available. For example, Cristea, *et al.* [9] attempted to use the date of registration (in terms of distance from the course start) of students to predict their completion of the course; Alamri, *et al.* [10] used the student's number of accesses and time spent per access in the first week of the course to predict their completion. However, only activity data, i.e., behavioural data, e.g., access to learning materials and discussion forums, were considered; whilst the demographical data, e.g., gender and educational level, might also be available at the start of the course, which might be considered as well to improve the prediction. Additionally, most existing studies, e.g., [11,12], classified students only into completers and non-completers (two categories), which might hide the differences amongst the students who completed a course, and the differences amongst the students who did not, even though a finer classification might be useful to understand why a student completes or drops out thus providing personalised interventions towards reducing the dropout rate as well as improving their participation and engagement.

Therefore, with the aim of moving towards bridging the gap, this study took into consideration both behavioural and demographical data. The objective was earlier prediction of finer classification of students in online learning especially within the context of MOOCs.

2 Related Work

Along the emergence of big data with the advances in computation, the areas of Learning Analytics (LA) and Educational Data Mining (EDM) have been rapidly developed in recent years, aiming at understanding how people learn online and improving the online learning process. While LA and EDM overlap with each other in similar attributes and goals, they are also different from each other in many aspects [13]. The former is stated as "the process of measuring and collecting data about learners and learning with the aim of improving teaching and learning practice" [14]; the latter is defined as "an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in" [15]. Both aim at improving the analysis of large-scale educational data to support practice in the educational context. In terms of their major differences, according to Siemens and Baker [13], in LA, leveraging human judgement is key, and automated discovery is a tool to accomplish this goal, while in EDM, automated discovery is key, and human judgment is a tool to accomplish this goal; LA has a stronger emphasis on understanding systems as a whole in full complexity, while EDM has a stronger emphasis on reducing components and analysing individual components and the relationships between them.

The main techniques and methods applied in LA and EDM include statistics, machine learning, and data mining, seeking usage patterns of learning resources including video lectures, forums, assessments, and so on, to compose useful models that can be smoothly adapted to educational data [16]. In particular, three techniques are often used in both LA and EDM: (1) prediction, to find a relationship between known and unknown data using simple statistical methods such as regression, non-linear statistics, and neural [17]; (2) clustering analysis, to create a collection of similar data objects within the same cluster [18]; and (3) relation mining, to classify various relationships that may occur between two or more variables [19].

While most studies, e.g. [20–22], focus on predicting completion and/or dropout rate, e.g., classifying students into completers and non-completers (two categories), we extend the predictive model and further classify students into four categories, including Withdrawn, Fail, Pass and Distinction. Besides, there are only a few similar studies, e.g. [9], [23], that tried predicting as early as possible student completion and dropout rate using limited data gathered. Our study also uses registration date as in previous studies [9] yet associated with also other parameters, as explained below in section 3, with the aim of producing a predictive model with better performance. Moreover, our predictive model aims to enhance the early predictive accuracy by introducing the BART (Bayesian Additive Regression Trees) model.

3 Method

3.1 Dataset

The dataset used in this study is the anonymised OULAD (Open University Learning Analytics Dataset)[1], which contains data about 7 courses and 32,593 registered students (55% males, 45% females), as well as their 10,655,280 interactions (clicks on webpages) with these 7 courses in the Virtual Learning Environment (VLE), operated by the Open University[2]. The dataset is in the format of 7 csv files, connected using unique identifiers including Student_ID, Assessment_ID, and Code_Module (ID of a course).

When joining the Open University for the first time, the students were directly prompted to complete an online form asking about their personal details such as gender and age. While using the VLE to study an online course, students' activity logs were generated, linked by unique Student IDs with timestamps, and recorded in the database. In total, these 7 courses provided 3,635 learning items, each of which was presented on a webpage in the VLE; there were 196 different assessments, and the students made 173,740 submissions. Interestingly, as Fig. 1 shows, out of 32,593 registered students, only 15,385 (42.78%) passed the courses, highlighting the fail/non-completion issue in MOOCs, which is in consistence with many reports, e.g., [4], [11], [24].

[1] https://analyse.kmi.open.ac.uk/open_dataset

[2] The OULAD dataset is released under CC-BY 4.0 licence

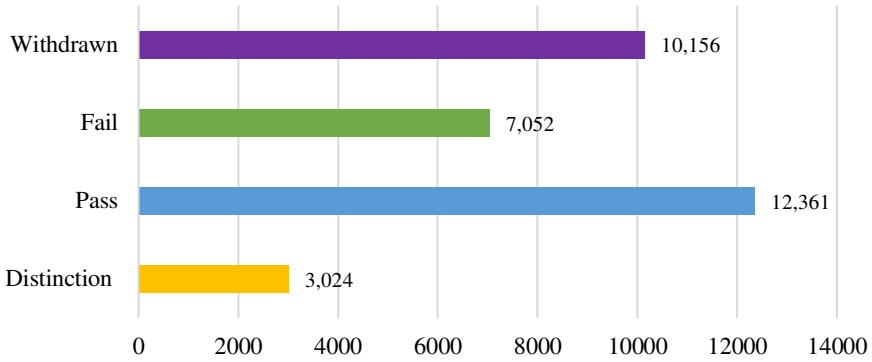

Fig. 1. Number of students in 4 categories: Withdrawn, Failed, Pass, Distinction

3.2 Study Settings and Data Preparation

The courses under study were organised in weekly learning units, each of which consisted of a collection of learning blocks that might contain one or a few steps. Steps were the fundamental learning items which might include articles, pictures, videos, and quizzes. Fig. 2 shows an example of the navigation page of a course, where a student might click one of the WEEK buttons to navigate to the weekly learning unit or click a step title to access a step page (learning item).

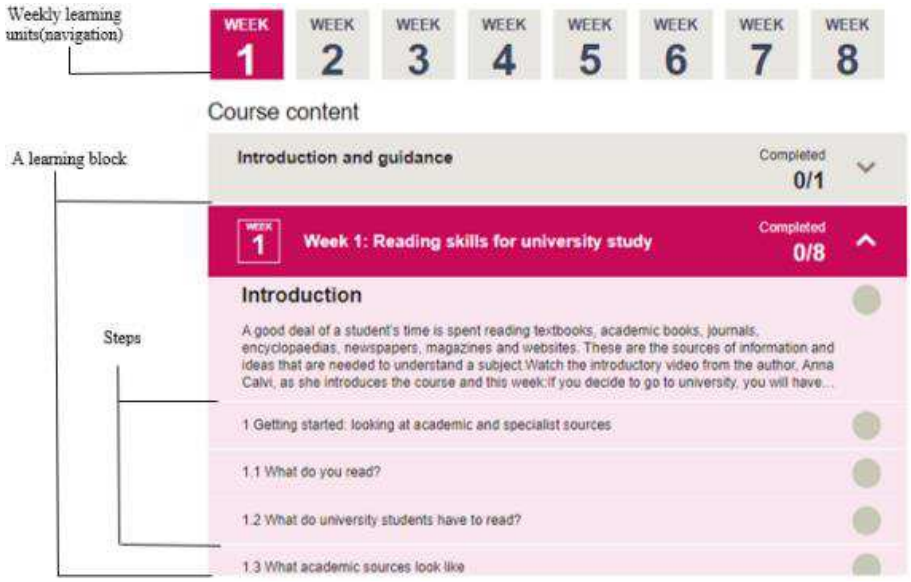

Fig. 2. Navigation page of a MOOC

It's worth mentioning that the courses in this study were "synchronous" – having official starting and finishing dates and running over an exact number of weeks [11]. In different courses, there were different numbers of assessments during a certain

period of time (week); additionally, at the end of each course, there was a final exam. Each course might change slightly, in different *runs* (i.e., years), the number of weekly learning units and steps, as well as assessment types (tutor marked assessment, computer marked assessment, and final exam). We used data from all 7 csv files as described in section 3.1. During a course, each student completed several assessments which had different weights summing up to 100%. We used the total number of clicks until a course started, for an *earlier prediction*. Each course had different durations and first assignment submission days, as shown in Table 1. We also converted the categorical variables including Educational Level and Age, into dichotomous variables.

Table 1. Information about MOOCs.

Course	1st assignment submission day	# of registered students	Year(run)
AAA	Day 19	748	2013 & 2014
BBB	Day 54	7,909	2013 & 2014
CCC	Day 18	4,434	2014
DDD	Day 23	6,272	2013 & 2014
EEE	Day 33	2,934	2013 & 2014
FFF	Day 19	7,762	2013 & 2014
GGG	Day 61	2,534	2013 & 2014

3.3 Analysis

For the analysis, seven variables were defined, as below.

- **First Assignment Mark**: the mark of a student's submission to the first assignment. On the StudentsAssessments csv file, it is called score.
- **Educational Level**: the highest level of education that a student has achieved; including 4 categories: Lower than A level, A level or equivalent, HE Qualification, and Post Graduate Qualification. On the StudentInfo csv file, it is represented as highest_education.
- **Clicks till Course Starts**: the number of clicks made by a student until a course started. Clicks are represented as sum_click on the studentVle csv file.
- **Registration Date**: the date of a student registered for a course, in terms of distance (the number of days) from the start of the course. On the studentRegistration csv file, it is represented as date_registration.
- **Age**: the band of a student's age (0-35, 35-55, >55). On the StudentInfo csv file, it is represented as age_band.
- **Disability**: whether a student has declared a disability. On the StudentInfo csv file, it is represented as disability.
- **Gender**: a student's self-reported gender (male/female). On the StudentInfo csv file, it is represented as gender.
- **Previous Attempts**: times that a student has failed a particular course. On the StudentInfo csv file, it is represented as num_of_pred_attempts.

We used the Pearson chi-square statistical hypothesis to test whether the output (Final Mark Classification) was dependent upon the categorical input variables (Educational level, Age, Gender, Disability), i.e., whether the input variables were relevant to the prediction tasks. The p-value was < 5%, which is within the acceptable range [25] , indicating that the categorical variables we used were relevant to the output. Moreover, to ensure that the variables were not only dependent upon the output, we also conducted Pearson's correlation tests to measure the strength of the association between the variables (results shown in Table 2), in terms of selecting variables which were not tightly related, in order to improve the predictive models' efficiency. Table 2 shows that the variables were correlated at a very low level showing that it was appropriate to use them as the input variables for our predictive models. The result of the two statistical tests shows that the selected variables fulfilled all the requirements in order to implement efficient and robust predictive models. The chosen variables for the resulting csv file used to train our learning algorithms included the First Assignment Mark, Educational Level, Clicks till Course Starts, Registration date, Age, and Gender. 70% of the data were used as the training data, and 30% as the test data. The majority of the algorithms we used relied on the default settings of the sklearn version 0.24.0, which can be found in the documentation for reference and reproduction[3]. The learning algorithms we used include Decision Tree, Random Forest, and BART, as they are known for their strong predictive power on binary classification problems.

Table 2. Pearson's Correlation test result

	Gender	Educational Level	Age	Previous Attempts	Disability	First Assig. Mark	Registration Date	Clicks till Course Start
Gender	1.00							
Educational Level	-0.03	1.00						
Age	0.02	0.15	1.00					
Previous Attempts	0.04	0.00	0.00	1.00				
Disability	0.04	-0.06	-0.02	0.04	1.00			
First Assignment Mark	-0.05	-0.01	0.04	-0.04	-0.04	1.00		
Registration Date	0.02	0.04	0.03	-0.02	-0.01	0.08	1.00	
Clicks till Course Starts	-0.10	0.03	0.12	-0.03	0.01	0.24	-0.07	1.00

Decision Tree is a supervised learning method which splits the population or sample into two or more homogeneous sets (or sub-populations) based on the most significant splitter/differentiator in input variables that predict the value of the target variable [26].

[3]https://pypi.org/project/scikit-learn/

Random Forest is a supervised learning algorithm that takes randomly selected data to build multiple decision trees merged together to generate more accurate and solid predictions. Specifically, Random Forest gets a prediction from each tree and selects the best solution using voting.

Bayesian Additive Regression Trees (BART), compared to Random Forest and Decision Tree, is the least used algorithm, so it is described in detailed. BART is a Bayesian version of tree ensemble methods where the estimation is given by the variable Y which is a sum of Bayesian CART trees [27]. We used the basic BART model which is shown in (1) below.

$$\Upsilon_\kappa = \sum_{j=1}^{m} g(x_\kappa; T_j, M_j) + \varepsilon_\kappa \quad (1)$$

In equation (1), T_j symbols the j^{th} decision tree j = 1 . . . m and M_j is a vector holding the terminal node parameters of T_j, while x_k is an n × p matrix of variables x, with $x_k = [x_{k1}, \ldots, x_{kp}]$, and $\varepsilon_k \sim N(0, \sigma^2)$, where σ^2 is the net variance (bias). In order to create a Bayesian model, we used a prior for the parameters, which in our case is the same as Chipman *et al.* [28] used:

$$P(T1, M1, T2, M2, ..., Tm, Mm, \sigma) = [\prod_j^m \{\prod_k^{bj} P(\mu_{\kappa j} | T_j)\} P(T_j)]P \quad (2)$$

From equation (2), we set distributions for the priors $\mu_{\kappa j}|T_j$, σ, and T_j which are $\mu_{\kappa j}|T_j \sim N(\mu_{\mu,} \sigma_\mu^2)$, $\sigma^2 \sim$ IG $(\frac{\nu}{2}, \frac{\nu\lambda}{2})$ and IG (α,β) respectively (α: the shape parameter, and β: the rate parameter). For ν, the default value is 3, and λ the value is determined in BART with the quantile set to 0.90.

To evaluate our predictive model's performance, we used the following four metrics.

- **Precision**: the ratio of the correctly predicted positive observations to the total predicted positive observations.
- **Recall**: the ratio of correctly predicted positive observations to all observations in the actual positive class.
- **F1-score** (3): the weighted average of Precision row and Recall row. Therefore, this score takes both false positives and false negatives into account.
- **Accuracy**: the most intuitive performance measure and it is simply a ratio of correctly predicted observations to the total observations.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

We used the "one-vs-rest" strategy, which fits a binary classifier for each class against all the rest of the classes, in particular – Withdrawn versus the rest, Fail versus the rest, Pass versus the rest, Distinction versus the rest. This allows binary classifiers (Decision Tree, Random Forest, BART (purely binary classifier)) to apply the already trained algorithm to an unseen sample x and predict the label y and calculating the performance of the algorithm with specific metrics. In our case, those metrics were Precision, Recall, F1-score and Accuracy. Specifically, we used precision and recall

metrics as those are better at characterising performance in the context of imbalance data (see Fig. 1).

4 Results and Discussions

Table 3 compares the performance of three similar tree-based algorithms that we used in the analysis, including Decision Tree, Random Forest, and BART. As mentioned in Section 3, we explored the BART model with the aim of improving our results and enhance the prediction accuracy. Interestingly, we found BART could give the optimum prediction accuracy on every "one-vs-rest" pair. Specifically, we achieved a relatively high accuracy of 81% for identifying students who might Withdraw from a course, 80% accuracy identifying students who might Fail, 69% accuracy identifying students who would get a Pass mark for the course, and 92% accuracy identifying students who might get a Distinction mark.

Table 3. Performance comparisons between three predictive models

	Metric	Decision Tree	Random Forest	BART
Withdrawn	Precision	0.65	0.75	0.81
	Recall	0.65	0.71	0.91
	F1	0.65	0.72	0.86
	Accuracy	0.69	0.78	0.81
Fail	Precision	0.68	0.69	0.79
	Recall	0.67	0.75	0.98
	F1	0.67	0.71	0.87
	Accuracy	0.67	0.76	0.80
Pass	Precision	0.63	0.65	0.72
	Recall	0.62	0.65	0.74
	F1	0.62	0.65	0.73
	Accuracy	0.63	0.65	0.69
Distinction	Precision	0.85	0.86	0.92
	Recall	0.84	0.90	0.98
	F1	0.85	0.87	0.96
	Accuracy	0.84	0.89	0.92

Table 4 shows the reason for a relatively low accuracy (yet, higher than Decision Tree and Random Forest), i.e., 69%, for the "Pass-vs-rest" pair classification, as the misclassified cases between the two classes is fairly high. As the Pass class is between the Fail class and the Distinction class, it seems that the algorithms tend to misclassify the Pass class as Fail or Distinction which is not happen.

Table 4. Confusion Matrix for Pass versus the rest

	Pass	**Rest**
Pass	1,582	1,529
Rest	1,436	3,879

Moreover, Fig. 3 shows the performance of the algorithms for the "Distinction-vs-rest" classification task, where we can observe the improved ability of the BART algorithm in comparison with Random Forest and Decision Tree algorithms to correctly classify the data.

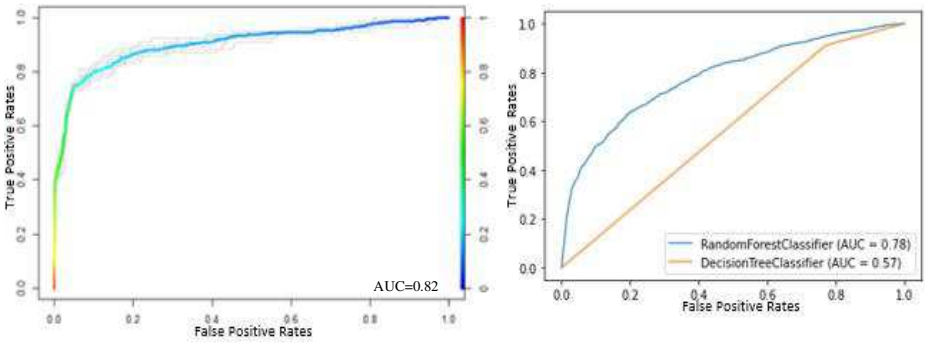

Fig. 3. BART (Left) Random Forest (Right) Decision Tree (Right) AUC graphs for Distinction versus the rest

Our results suggest that combining demographical data (such as educational level, gender, age, and disability) and behavioural data (such as student's daily activity (clicks), the number of previous attempts in a course, first assignment mark, and registration date) can produce a predictive model with good performance.

The results obtained are worthy of discussion - as we observe that among the tree-based machine learning algorithms we used, the BART outperforms the others. To begin with, our results show that BART produced the optimal predictive accuracy for every "one-vs-rest" pair (i.e., Withdrawn, Fail, Pass, Distinction, respectively, with the rest of the classes). Our model could predict the final result classification (Withdrawn, Fail, Pass, Distinction), so the lecturers, after the first assignment, can use it to identify who is more likely to Fail, Pass, etc., thus being able to provide early interventions to these students, with tailored reminders, as the students were classified into finer-grained categories (comparing to other methods that classified them into only two categories – completers and non-completers).

It is very important to highlight the strong predictive power of the number of clicks (resource, glossary, URL, forum, homepage, etc) on the VLE, which we should aim to raise in order to improve students' performance. Fig. 4 shows that students who failed (green dots) exhibit significantly a smaller number of clicks on the VLE compared to those with a pass (blue dots) or a distinction (yellow dots) mark. This suggests that high scores are associated with more frequent access to the VLE, and that, in order to have a better result of the course, students should be using the VLE more often.

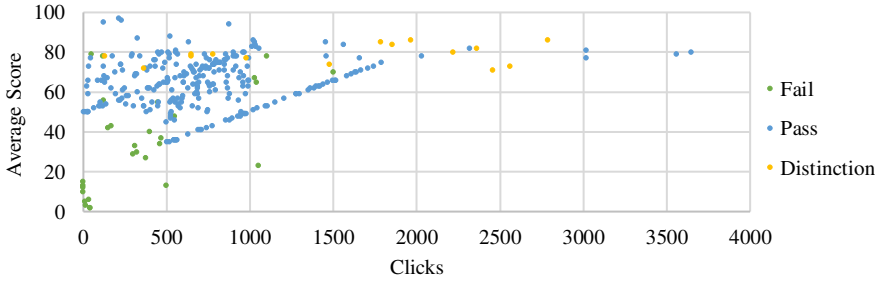

Fig. 4. Relationship between the number of clicks and the overall student outcome

5 Conclusions

In summary, this paper presents the results of a study aiming to discover whether it is possible to predict and identify, as early as possible, which students might withdraw from a course, and, possibly, make earlier interventions to reduce their withdrawal or failure, and to improve students' final marks. This is different from most previous studies that analysed the data after the completion of the whole course which is not very useful for the **current** students. To produce and validate the predictive models, we have examined 8 independent variables in total, including both demographical variables (*Educational Level*, *Gender*, *Age*, *Disability*), and behavioural variables (*Registration Date*, *Clicks until Course Starts*, *First Assignment Score*, and *Previous Attempts* on Open University's (OU) VLE). This is different from most previous studies where only behavioural variables are included.

The main limitation, however, was the strict scope of the dataset. The daily interaction with the VLE, i.e., clicks, plays an important role but the virtual learning system (VLE) is not integral. For example, the results of the final written exams were not included in the csv files. Besides, on the independent variable Cli*cks till Course Starts,* we could not take into consideration the students' educationally relevant discussions outside of the OU's VLE or the private discussion forums, and it is worth noting that not all learning behaviour could be fully captured through online platforms.

Future work may include investigating and validating efficient strategies for the use of the proposed predictive model. For example, it could be used in 3 different stages of a MOOC. Firstly, use the model to identify, as early as possible, the students who are likely to withdraw. For example, in order to keep the student remaining in a course, the lecturer could send personalised messages reinforcing the usefulness and objectives of the course. Secondly, after a couple of weeks, when more data is collected such as the second assignment mark, the lecturer could use the model to identify students who might fail with improved accuracy and provide them with necessary supports. Finally, at the final stage of the MOOC (previous assignments marks could have been added to the model as an additional input) before the final examination, the model can be used to identify the students with Pass or Distinction marks and provide the lecturer with a precise overview of the students' benchmarks. Importantly, the first assignment mark is suggested to be a very strong predictor of students' performance. Thus, the lecturer is recommended to periodically send students reminders with evidence, to emphasise the importance of participation and engagement to be successful in a course.

References

1. 'By The Numbers: MOOCs in 2020 — Class Central', *The Report by Class Central*, Nov. 30, 2020. https://www.classcentral.com/report/mooc-stats-2020/ (accessed Jan. 04, 2021).
2. 'Study offers data to show MOOCs didn't achieve their goals | Inside Higher Ed'. https://www.insidehighered.com/digital-learning/article/2019/01/16/study-offers-data-show-moocs-didnt-achieve-their-goals (accessed Jan. 04, 2021).
3. M. G. Gomez-Zermeno and L. A. D. L. Garza, 'RESEARCH ANALYSIS ON MOOC COURSE DROPOUT AND RETENTION RATES', 2016, doi: 10.17718/TOJDE.23429.
4. F. Dalipi, A. S. Imran, and Z. Kastrati, 'MOOC dropout prediction using machine learning techniques: Review and research challenges', in *2018 IEEE Global Engineering Education Conference (EDUCON)*, Apr. 2018, pp. 1007–1014, doi: 10.1109/EDUCON.2018.8363340.
5. I. Borrella, S. Caballero-Caballero, and E. Ponce-Cueto, 'Predict and Intervene: Addressing the Dropout Problem in a MOOC-based Program', in *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*, Chicago IL USA, Jun. 2019, pp. 1–9, doi: 10.1145/3330430.3333634.
6. M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart, 'Predicting MOOC Dropout over Weeks Using Machine Learning Methods', in *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, Doha, Qatar, Oct. 2014, pp. 60–65, doi: 10.3115/v1/W14-4111.
7. J. Liang, C. Li, and L. Zheng, 'Machine learning application in MOOCs: Dropout prediction', in *2016 11th International Conference on Computer Science Education (ICCSE)*, Aug. 2016, pp. 52–57, doi: 10.1109/ICCSE.2016.7581554.
8. J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley, 'MOOC Dropout Prediction: How to Measure Accuracy?', in *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, Cambridge Massachusetts USA, Apr. 2017, pp. 161–164, doi: 10.1145/3051457.3053974.
9. A. Cristea, A. Alamri, C. Stewart, M. Alshehri, and L. Shi, 'Earliest Predictor of Dropout in MOOCs: A Longitudinal Study of FutureLearn Courses Mizue Kayama', Aug. 2018.
10. A. Alamri *et al.*, 'Predicting MOOCs Dropout Using Only Two Easily Obtainable Features from the First Week's Activities', in *Intelligent Tutoring Systems*, vol. 11528, A. Coy, Y. Hayashi, and M. Chang, Eds. Cham: Springer International Publishing, 2019, pp. 163–173.
11. Y. Wang and R. Baker, 'Content or platform: Why do students complete MOOCs?', vol. 11, no. 1, p. 14, 2015.
12. L. Uden, J. Sinclair, Y.-H. Tao, and D. Liberona, Eds., *Learning Technology for Education in Cloud. MOOC and Big Data*, vol. 446. Cham: Springer International Publishing, 2014.

13. E. Baran, 'Siemens & Baker, "Learning Analytics and Educational Data Mining: Towards Communication and Collaboration"', in *Learning Environments Design Reading Series*, .
14. 'learning analytics | Advance HE'. https://www.advance-he.ac.uk/knowledge-hub/learning-analytics (accessed Mar. 29, 2021).
15. 'educationaldatamining.org'. https://educationaldatamining.org/ (accessed Mar. 29, 2021).
16. L. C. Liñán and Á. A. J. Pérez, 'Mineria de dades educatives i anàlisi de dades de l'aprenentatge: diferències, semblances i evolució en el temps', *RUSC. Universities and Knowledge Society Journal*, vol. 12, no. 3, Art. no. 3, Jul. 2015, doi: 10.7238/rusc.v12i3.2515.
17. C. D. Madigan, A. J. Daley, E. Kabir, P. Aveyard, and W. Brown, 'Cluster analysis of behavioural weight management strategies and associations with weight change in young women: a longitudinal analysis', *Int J Obes*, vol. 39, no. 11, pp. 1601–1606, Nov. 2015, doi: 10.1038/ijo.2015.116.
18. '4 - Prediction.pdf'. Accessed: Mar. 29, 2021. [Online]. Available: http://www.cs.stir.ac.uk/courses/ITNP60/lectures/1%20Data%20Mining/4%20-%20Prediction.pdf.
19. I. P. Klapaftis, S. Pandey, and S. Manandhar, 'Graph-Based Relation Mining', in *Multimedia Communications, Services and Security*, Berlin, Heidelberg, 2011, pp. 100–112, doi: 10.1007/978-3-642-21512-4_12.
20. P. J. Guo and K. Reinecke, 'Demographic differences in how students navigate through MOOCs', in *Proceedings of the first ACM conference on Learning @ scale conference*, New York, NY, USA, Mar. 2014, pp. 21–30, doi: 10.1145/2556325.2566247.
21. L. Shi and A. Cristea, 'Demographic Indicators Influencing Learning Activities in MOOCs: Learning Analytics of FutureLearn Courses', Aug. 2018.
22. J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley, 'Delving Deeper into MOOC Student Dropout Prediction', *arXiv:1702.06404 [cs]*, Feb. 2017, Accessed: Jan. 28, 2021. [Online]. Available: http://arxiv.org/abs/1702.06404.
23. C. G. Brinton and M. Chiang, 'MOOC performance prediction via clickstream data and social learning networks', in *2015 IEEE Conference on Computer Communications (INFOCOM)*, Apr. 2015, pp. 2299–2307, doi: 10.1109/INFOCOM.2015.7218617.
24. T. R. Liyanagunawardena and S. A. Williams, 'Dropout: MOOC Participants' Perspective', p. 8.
25. S. D. Bolboacă, L. Jäntschi, A. F. Sestraş, R. E. Sestraş, and D. C. Pamfil, 'Pearson-Fisher Chi-Square Statistic Revisited', *Information*, vol. 2, no. 3, Art. no. 3, Sep. 2011, doi: 10.3390/info2030528.
26. '1.10. Decision Trees — scikit-learn 0.24.1 documentation'. https://scikit-learn.org/stable/modules/tree.html (accessed Mar. 29, 2021).
27. H. A. Chipman, E. I. George, and R. E. McCulloch, 'BART: Bayesian additive regression trees', *arXiv:0806.3286 [stat]*, Oct. 2010, doi: 10.1214/09-AOAS285.