Machine Learning Approaches to Predict Learning Outcomes in Online Courses

Charles L. Green
Electrical Engineering and Computer Science
Pennsylvania State University
State College, PA, United States of America
CharlesLGreen14@gmail.com

Dr Jiangtao Huang
Computer Science and Engineering
Michigan State University
East Lansing, MI, United States of America
huangj90@msu.edu

Yao Ma
Computer Science and Engineering
Michigan State University
East Lansing, MI, United States of America
mayao4@msu.edu

Dr Jiliang Tang
Computer Science and Engineering
Michigan State University
East Lansing, MI, United States of America
tangjili@msu.edu

***Abstract*—Improving the learning experience can be done with the use of Learning Analytics. Based on the Open University Learning Analytics dataset, in 2013 and 2014, 47.25% of the people in the selected modules passed while 21.62% failed and 31.13% drop the class. This could be for a multitude of reasons age, gender, location or activity on the online course. Finding which one of the many influences on their grades are the most impactful on the final outcome of the class can help raise the number of people passing and passing with distinction, while also lowering the drop and fail rate of classes. The goal is to test different machine learning algorithms and to compare their performances for this dataset. We will also use the data obtained from machine learning to help find which influences are most important in predicting the final outcome of a class. Demographics about 32,593 students, their assessment results and the logs of their interactions on the online courses will be used to test machine learning algorithms such as Decision Trees and Support Vector Machines, amongst other Machine Learning Algorithms. With the analysis of the test, a strong correlation between 3 features and the final outcome of a course was found, the top feature being Percentage of Assessments Taken. With the algorithms tested, Gradient Boosted Decision Trees and Random Forest achieved the highest accuracy with Naive Bayes having the highest efficiency. Conversely, Support Vector Machines achieved the worst in both of these categories.**

I. Introduction

Massive open online courses (MOOCs) got its name in 2008 by David Cormier and George Siemens. They used this name to describe their 12-week course on connectivism at the University of Manitoba, Canada. Massive refers to the large scale of the course while the Open Online courses part describes the free/inexpensive access to the online learning material. [5] These courses tend to collect records of users interactions with the online course from things such as clicks on certain videos or in total, to demographics and possible learning behaviours. [1] This record of users interactions is called Learning Analytics, which can be used to analyse these behaviours that can help improve these MOOCs.

MOOCs have been growing at a rapid rate of recently. As of 2018, there have been 101 million people who have signed up for an MOOC and in 2018 alone there were 20 million new learners who have signed up for one or more MOOC. [9] Despite these high rates of signing up for a MOOC, the drop out rates is still significantly high than the in-class counterparts to these classes. [6] While this could be for many reasons, Learning Analytics on these MOOCs can help determine some of the reason why these rates are so high, and what can be done to lower these rates and raise the passing rates of the online courses.

MOOCs have been used by many to help adapt to the ever-changing society around them and may become essential for society in the future so everyone can continue growing their knowledge in their field of study. With the necessity of these courses, an improvement upon the rates must happen. These reasons can go from starting late too little expectations to the length of the course. [10] Previous research has been done on the Learning Analytics of many different online courses. They have used methods such as prediction, clustering, statistics, machine learning and many more methods. Now while there has been 200 publication on learning analytics of MOOCs as of 2017, very few have been used with the method of machine learning (11) [5]. Machine learning is the scientific study of algorithms and statistical models used to perform a specific task without explicit instructions. This can help find these hidden insights of the data and the weights of all possible influences on the final outcome of the online course, such as passing with distinction, passing, failing and withdrawing.

From some previous publications on studying learning analytics with machine learning, some may use just a binary outcome of if the user finishes the course with a pass and a certificate or a fail [3] while others may look for if the user drops the course or not and if so, when they dropped the course. [7] This has some useful qualities in their individual ways but does not fully solve all the possible outcomes. The solution that has taken place in this project is to first use

many different machine learning to test in the accuracy in predicting the outcomes of passing with distinction, passing, failing and withdrawing. Second will be to use all the machine learning algorithms to test the weights or significance of each possible influence on the final outcome. This solution will help test more possible outcomes some previous test while also figure out the specific influences why a person may have those outcomes and the importance of these influences. With the Open University Learning Analytics dataset, the users from these datasets can be used with different machine learning algorithms to help figure out the reasons for the previously states final outcomes of the course. [2] This can help improve these MOOCs if their structure to make it easier for a student pass and learn from these courses and can help students figure out better ways to pass these course with relative ease.

II. Methodology

A. Data Organization

The data that will be used is from the Open University Learning Analytics dataset (OULAD). This data has been comprised of data that can be used to improve the learning experience by providing informed guidance and to help improve the learning materials in the future. This data contains 7 CSVs about 22 courses from 2013-2014, 32,593 students, their assessment results, and logs of their interactions with the VLE with daily summaries of a student's clicks. A CSV is a tabular way of exporting data separated by commas and newlines. These CSVs are all structured in different ways must be organized prior to the machine learning testing. All information from the CSVs will be organized into a single CSV in the format of the 32,593 students. The data will originate student info CSV because of the original 32,592 student format it currently has. Information from student registration shall be added next to add the registration and unregistration date. The length of each course will be used with the assessments to help simplify the data from the student Virtual Learning Environment (VLE) from the 10,655,280 entries to 32,593. This will be done by adding columns such as average interactions per day the user interacted with the course and a percentage of days in the course the user interacted. First interaction, and last interaction will be added to the CSV to help with testing. Finally, with the CSV on the assessments for each student, their total number of assessments taken and an average grade prior to final will be added. In the end, there will be 18 different features that can be used to help the Machine Learning Algorithms find the final outcome. [2]

With all 7 CSVs impactful data being simplified into one CSV, the categories chosen to be used will be determined by which should have no impact of the final outcome and which may cause data leakage with the machine learning algorithms. Data leakage is when information outside the training dataset is used to create the model, can invalidate predictive models. With all the data provided by Open University Learning Analytics Dataset (OULAD), a variable that should have no influence on the final outcome should be user IDs so that will be dropped in testing the machine learning algorithms. A variable that may cause data leakage may be date unregistered. This would most likely cause data leakage in the testing due to 100 percent of the people who have a value in the unregistered category would have a 100 percent drop rate of the course. This will skew the results and make the machine learning test invalid so a category that may cause data leakage such as date unregistered shall not be used in testing. Finally, for highly similar categories, it may be chosen which ones may have to be removed. For example, if a user who lives in a certain area tend to fail/drop more often, the final conclusion may come to that area needing more help in education or motivation. However, that area may also be an impoverished area, which makes time to take the MOOC to be less prevalent and that could be a reason. Something like that will have to be taken into consideration while testing so a situation like this does not occur.

B. Machine Learning Testing

There will be various Machine Learning Algorithms that will be tested to try to get the results. They are Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), Gradient Boosted Decision Trees (GBDT), K-Nearest Neighbors (KNN) and Neural Networks (NN). These are described in table 1, model description. In this study, it will be investigated which algorithm is most accurate in predicting the final outcome, which will be displayed in percentages, and the average weight of each possible influence.

To get these results, the Python coding language will be used with the Anaconda Navigator to open up the Jupyter Notebook IDE. This IDE will have all the required libraries such as sci-kit learn to use the machine learning algorithms necessary. The previously sorted data will be imported into the IDE to be used for machine learning. With the dataset of 32,593 users, it will be separated into training and test data. 70 percent of this data will be used in training the algorithm while 30 percent of this data will be used to test the algorithm. The training data will then be separated by k-fold cross-validation to attempt to overcome overfitting from the original data with k=5. With the sci-kit learn library in python, it makes the k-fold cross validation become stratified. This makes the k parts of the data be keeping the same percentages of samples of each target class in each k set, and trains data with k-1 training and 1 testing. As the machine learning models are training and testing the data, all parameters will be optimized for the best results. These machine learning algorithms will be tested 3 times with different random state variables to allow consistent but different results that will be averaged to get more accurate results. Then each variables feature importance will be found with the ExtraTrees classifier in the sci-kit learn library. This class uses randomized decision trees of various sub-samples of the dataset and uses averaging to improve the predictive accuracy of the feature importance of each variable/influence. It will also be tested twice with all the variables and then the highest weighted half of the variables to compare those results. Finally, all algorithms will be tested again with binary results of pass or not pass outcome to compare the 2 results. With

Model	Description	Architecture	Type	Algorithm
DT	Decision Tree	Recursive Partition Decision rules	Non Linear	CART algorithm
RF	Random Forest	Ensemble DT	Non Linear	Random Subset Features Bootstrap
SVM	Support Vector Machine	Hyperplane kernel trick	Non Linear	Quadratic Optimization
NB	Nave Bayes	Bayesian Decision Rule	Linear	Maximum Likelihood Estimation
GBDT	Gradient Boosted Decision Trees	Ensemble DT	Non Linear	CART algorithm
KNN	K-Nearest Neighbors	Instance based learning	Non Linear	KDTree
NN	Neural Networks	Multi-layer Perceptron	Non Linear	Backpropagation

TABLE I
MACHINE LEARNING ALGORITHMS

these results, it can be analyzed on which machine learning methods are best in predicting outcomes on MOOCs and it will be determined which of the variables that are most influential in the final outcome of the course.

III. RESULTS

A. Experiment Setup

In the Juypter Notebook, the libraries of numpy, scipy, pandas, matplotlib, seaborn, sklearn was imported to the file for the use of the ML algorithms and for the use of analysis of the algorithms are working. In this case study, the data is segmented into 70 percent training data and 30 percent testing data. This leaves 22815 users in the training data and 9778 users in the testing data. There will be 3 random state variables of 0, 1, and 10 to allow for different results in each test of the algorithm, that will be averaged out for a more accurate result. The standard deviation of the result will be with the result multiplied by 2 to allow for greater variance due to the unlimited amount of random state variables that may lead to slightly different results. The most important features will also be calculated for this test with the Extra Trees Classifier in the sklearn library. This uses an assortment randomized DT on various sub-samples of the dataset to determine the feature's importance. This will be done with 1000 estimators with the random state set to 0. With the top half of the features, all algorithms will be tested again to see for a difference in the accuracy and speed. Stratified k-fold cross validation was applied during the test to overcome over fitting the data. The subsets are set to 5 equal sized subsets of data, keeping same percentage of samples of each target class in each k set. There will be 4-fold subsets used as the training set and the remaining subset used as a test sample.

For the sklearn library, most parameters for the ML algorithm but some has been changed to get optimal results. For the Gradient Boosted Decision Trees, all settings were set at the default values except the number of estimators were set to 50 trees. For the Random Forest Algorithm, the maximum features were set to 9 and the max depth was set to 10, with the rest of the settings staying default. In the Decision Tree Algorithm, the minimum samples leaf was set to 2, the maximum depth was set to 10 and the maximum leaf nodes was set to 70. For K-Nearest Neighbors, the k-neighbors was set to 30 with the power parameter for the Minkowski metric set to p=1, which is equivalent to Manhattan distance. Manhattan distance being equivalent to the sum of the horizontal and vertical distances between points on a grid. For the Neural Network Algorithm, there were 2 hidden layers set with each layer having a size of 8 neurons. The solver for this algorithm was set to 'adam' which refers to a stochastic gradient-based optimizer. For the Naive Bayes algorithm and the Support Vector Machine, all settings were kept at the default settings that the sklearn library provides. All parameters not mentioned above was kept at the default value sklearn has them set as.

B. Result Evaluation

This section contains the results of the lab in terms of accuracy which is measured in the form of (TP)/(TP+FN) Standard Deviation 2. TP being the True Positive and the FN being the False Negative. Standard Deviation is multiplied by 2 due to the unlimited amount of random state variables which may all lead to slightly different results. The multiplication by 2 accounts for the total error that may occur. Throughout the test, the highest stated Standard Deviation was taken a multiplied by 2. For the efficiency, the timeit library was imported to the code to keep track of the run time for the code. The shortest amount of time it took the code to run was taken for the results. As most computers code like this can be ran on are different in speeds, the times can be used to compare to each other to see which ones are faster than the others. For the Feature Importance, all 18 features were set into a value of 1 and based on the Extra Trees feature importance tester, all values were given a value and if added up equals 1. The results of this can be found in figure 1. The results will be seperated in terms of the tests with all features and test with the top 9 or half of the features, denoted as test 1 and test 2 respectively. Also the Binary Results, listed in table 3, will be listed and compared to the results containing the final outcome of pass, fail, pass with distinction and withdraw, which can be found in table 2. Both these tables results are based on confusion matrix matrices used with the sklearn library. The simulation results from test 1, associates all data set features, found in figure 1, yields slightly higher results than in test 2 which associates the top 9 features. The GBDT algorithm achieved the highest accuracy of 0.750 ± 0.018 in the first test and 0.739 ± 0.013 in the second test. Although these achieved the highest accuracy, the efficiency for this algorithm was the second slowest out of all the tested algorithms with a speed of 11.41 seconds on the first test and 7.760 seconds on the second test. This efficiency of this algorithm makes it seem less suitable then some of

the other ones. The RF algorithm was second in its accuracy with 0.744 $\pm$ 0.015 in the first test and 0.732 $\pm$ 0.011 in the second test. The efficiency of this algorithm is approximately 10 times faster than GBDT in test 1 with a 1.198 seconds and approximately 5 times faster in test 2 with 1.733 seconds. Test 2 was slower than test 1 in this algorithm unlike the majority of the other algorithms, even though less features was used. DT was third in accuracy with 0.735 $\pm$ 0.013 and an efficiency of 0.281 seconds which is approximately 4 times faster than RF. The accuracy steadily decreased as the other ML algorithms were tested with in Test 1 NN, KNN, and NB having a 0.710 $\pm$ 0.021, 0.716 $\pm$ 0.014 and 0.655 $\pm$ 0.021 accuracy respectively. The Efficiency of these Algorithms varied with 5.877 seconds, 2.414 seconds, and 0.041 seconds respectively. With test 2, the accuracy lowered slightly with NN, KNN, and NB with 0.703 $\pm$ 0.027, 0.713 $\pm$ 0.015 and 0.654 $\pm$ 0.013 while the efficiency got worse in NN and better in KNN and NB with an efficiency of 9.955 seconds, 1.225 seconds, 0.032 seconds respectively. For the SVM algorithm, it was the worse in both accuracy and efficiency with an accuracy of 0.451 $\pm$ 0.009 and 0.477 $\pm$ 0.011 in test 1 and 2 and efficiency of 219.378 seconds and 162.927 seconds. Based on these results, the top 3 test of GBDT, RF and DT are the best when it comes to accurately predicting the final outcomes of MOOCs with GBDT, and RF trading away some efficiency for the sake of being slightly more accurate. Also when using only the top weighted features, these algorithms stay in the top 3 but RF becomes slower while the others become a little more efficient.

Getting results with 4 variables, (Pass, Pass with Distinction, Fail and Withdraw), the accuracy seems lower than binary results of Pass and Not Pass. The worst algorithm is still SVM with an Accuracy of 0.644 $\pm$ 0.009 and 0.666 $\pm$ 0.014 and an efficiency of 170.283 seconds and 134.988 seconds in test 1 and 2 respectively. In the rest of the algorithms, the accuracy and efficiency both improve. The top 3 most accurate algorithms are still GBDT, RF and DT but the accuracy improved by approximately 20 percent with an 0.940 $\pm$ 0.011, 0.939 $\pm$ 0.011, and 0.935 $\pm$ 0.012 in test 1. In test 2 with the top featured variables, the accuracy insignificantly lowers if it lowers at all compared to test 1 with GBDT, RF and DT having accuracy's of 0.939 $\pm$ 0.012, 0.937 $\pm$ 0.012, and 0.935 $\pm$ 0.012 respectively. The efficiency's of these algorithms improved little in all algorithms except for GBDT which got approximately 4 times faster than the original results of 4 variables with efficiency's of 2.689 and 1.737 in test 1 and 2 respectively. These high boost in accuracy's of the algorithms between the non binary and binary results could have happened for a multitude of reasons. The one that may be most likely however is how Pass and Pass with Distinction have highly similar data with minor differences as well as Withdraw and Fail have similar data. This most likely confused the algorithms due to all the similarities in features. With these more accurate results, it may help in predicting the final outcome more in just getting a user to simply pass, but with the 4 outcomes, it can help pinpoint what a user can do to help improve there grade from pass to passing with distinction.

ML Alg.	Test 1 Acc.	Test 1 Eff.	Test 2 Acc.	Test 2 Eff.
GBDT	0.750 $\pm$ 0.018	11.41	0.739 $\pm$ 0.013	7.760
RF	0.744 $\pm$ 0.015	1.198	0.732 $\pm$ 0.011	1.733
DT	0.735 $\pm$ 0.013	0.281	0.725 $\pm$ 0.016	0.223
NN	0.710 $\pm$ 0.021	5.877	0.703 $\pm$ 0.027	9.955
KNN	0.716 $\pm$ 0.014	2.413	0.713 $\pm$ 0.015	1.225
NB	0.655 $\pm$ 0.021	0.041	0.654 $\pm$ 0.013	0.032
SVM	0.451 $\pm$ 0.009	219.378	0.477 $\pm$ 0.011	162.927

TABLE II
ML TEST RESULTS WITH 4 FINAL OUTCOMES

ML Alg.	Test 1 Acc.	Test 1 Eff.	Test 2 Acc.	Test 2 Eff.
GBDT	0.940 $\pm$ 0.011	2.689	0.939 $\pm$ 0.012	1.737
RF	0.939 $\pm$ 0.011	0.971	0.937 $\pm$ 0.012	1.353
DT	0.935 $\pm$ 0.010	0.209	0.935 $\pm$ 0.012	0.168
NN	0.932 $\pm$ 0.011	3.249	0.929 $\pm$ 0.012	7.787
KNN	0.932 $\pm$ 0.012	2.316	0.934 $\pm$ 0.012	1.17
NB	0.917 $\pm$ 0.008	0.035	0.918 $\pm$ 0.008	0.029
SVM	0.644 $\pm$ 0.009	170.283	0.666 $\pm$ 0.014	134.988

TABLE III
ML BINARY TEST RESULTS WITH PASS/ NOT PASS FINAL OUTCOMES

With these predictions it may help in getting a student who on track of failing to change some habits or help improve the program to make it so they are more likely to pass.

With the Feature Importance test, the majority of the features were expected to be ranked where they were. The majority of the demographics and the features which are determined prior to taking the course such as gender, age, and disability had a low importance in predicting the final outcome. Features such as Percentage of Assessments Taken, Average Grade Prior to Final and Last Interaction had most importance and similar importance of of all the features with 0.177, 0.163, and 0.143 importance respectively. The only surprising part of the Feature importance is how the Highest Education feature was in the bottom half of the important variables. It would be expected that how much education you had prior to taking the class would hold a bit more importance but with the variables that are listed to have more importance, it is understandable where it is ranked in the list.

C. Discussion

The experiments in this study were aimed to predict the final outcome in MOOCs. Pre-processing methods to make sure all data was organized in 1 CSV and each user had their own line of data. In this paper, four types of experiments have been conducted. In the first 2 sets of experiments, the final

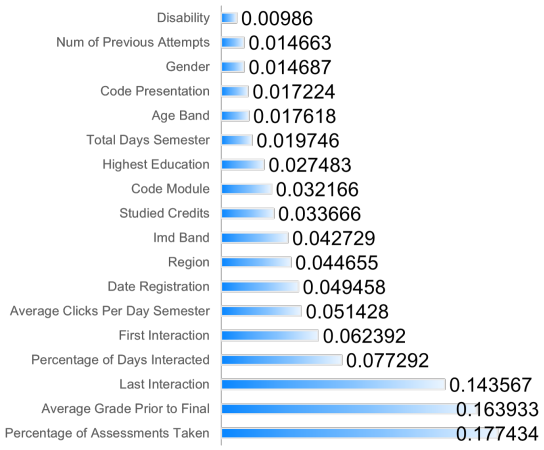

Fig. 1. Feature Importance of Each Influence/Variable

outcomes of the courses were Pass, Pass with Distinction, Fail, and Withdraw, with 1 experiment using all features with the other using only the top half of the features. The final 2 sets of experiments used the binary set of outcomes of Pass and Not Pass with the first test using all features and the second using the top half the the features. All these test were compared to each other, observing the similarities and differences in accuracy and efficiency of each algorithm. For the non binary test, the accuracy ranged from 0.654 to 0.750 with the an outlier with the SVM which had under 0.500 accuracy. With the binary results, the accuracy ranged from 0.917 to 0.939 with an outlier with the SVM which had an accuracy under 0.700. GBDT, RF and DT had the highest accuracy's but varied in efficiency's. The efficiency's in the non binary test ranged from 0.032 seconds to 11.410 seconds with SVM having taking over 2 minutes in both test. In the binary test, the efficiency ranged from 0.029 seconds to 2.689 seconds with SVM taking over 2 minutes. The most efficient models in all test were NB and DT taking under a second in all test. The Binary results compared to the Non Binary results were all more accurate and efficient compared to their counterparts. Out of all 18 features, only 3 had over 0.1 in importance Percentage of Assessments Taken, Average Grade Prior to Final and Last Interaction being the most important features.

IV. Conclusion and Future Work

This study was taken to examine the effectiveness of machine learning approaches in the prediction of final outcomes within MOOCs. Effectiveness was measured in terms of accuracy and efficiency. In this work, four set experiments have been applied. In the first sets, non binary results were used with all features, with the second set using the top half of the features. In the last sets, binary results were used with all feeatures, and finally with the top half of features with the last test. Various machine-learning approaches have been applied over all experiments to predict learning outcomes relating to the Open University Learning Analytics dataset.

The simulation results in the experiments showed how while GBDT had the highest accuracy in all test, RF and DT were close behind in accuracy while having a significantly higher efficiency. NB had the highest efficiency but also the second lowest accuracy. SVM had the worst accuracy and efficiency of all of the algorithms. The Binary results all had a significantly high accuracy in predicting the final outcome and a better efficiency then their non binary counter parts. The results show how that ML is a viable approach to figuring out the final outcomes, but when using more specific outcomes the accuracy suffers. Also the average run time of the results using more features on average has a worse efficiency with a few outliers. With the features given, it was also seen that the features related to demographics and predetermined features had less importance in predicting the final outcome. Future work can be done to investigate other features that may occur within MOOCs that may effect the learning outcome. Also features that cannot be explicitly tracked from the MOOC course itself but with a survey can be used. For example, learning about a users emotional state or if an event is happening in a users life can help effect the outcome of a course. If this data can be obtained, these features may be important in future testing, allowing the non binary outcomes to have higher accuracy and be closer to their binary counterparts. Also in the future, different algorithms on the same or different datasets to see if different sized data. These test can be used to help improve future MOOC course and improve the chances passing these courses.

Acknowledgment

I would like to thank the following people for their help and mentoring on this project. Yao Ma, Dr Huang, Dr Tang and Mr. Steven Thomas. I would also like to thank the Michigan State University Summer Research Opportunities Program (MSU SROP) and its faculty and facilitators along with the Millennium Scholar Program.

References

[1] J. Qiu et al., Modeling and Predicting Learning Behavior in MOOCs, Proc. Ninth ACM Int. Conf. Web Search Data Min., pp. 93102, 2016.

[2] Kuzilek, J., Hlosta, M. Zdrahal, Z. figshare https://doi.org/10.6084/m9.figshare.5081998.v1 (2017).

[3] Al-Shabandar, Raghad, et al. Machine Learning Approaches to Predict Learning Outcomes in Massive Open Online Courses. 2017 International Joint Conference on Neural Networks (IJCNN), 2017, doi:10.1109/ijcnn.2017.7965922.

[4] Baturay, Meltem Huri. An Overview of the World of MOOCs. Procedia - Social and Behavioral Sciences, vol. 174, 2015, pp. 427433., doi:10.1016/j.sbspro.2015.01.685.

[5] Khalil, Mohammad. MOOCs Completion Rates and Possible Methods to Improve Retention - A Literature Review. Graz University of Technology, 2017.

[6] Kashyap, Avinash, and Ashalatha Nayak. Different Machine Learning Models to Predict Dropouts in MOOCs. 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, doi:10.1109/icacci.2018.8554547.

[7] Dalipi, Fisnik, et al. MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges. 2018 IEEE Global Engineering Education Conference (EDUCON), 2018, doi:10.1109/educon.2018.8363340.

[8] Kloft, Marius, et al. Predicting MOOC Dropout over Weeks Using Machine Learning Methods. Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, 2014, doi:10.3115/v1/w14-4111.

[9] By The Numbers: MOOCs in 2018 - Class Central. Class Central's MOOC Report, 21 Mar. 2019, www.classcentral.com/report/mooc-stats-2018/.

[10] Onah, D.F.O., et al. DROPOUT RATES OF MASSIVE OPEN ONLINE COURSES: BEHAVIOURAL PATTERNS. 2014, p. 10.