Applying Machine Learning to Predict Whether Learners Will Start a MOOC After Initial Registration

Theodor Panagiotakopoulos[1,2](✉), Sotiris Kotsiantis[3], Spiros Borotis[1], Fotis Lazarinis[1], and Achilles Kameas[1]

[1] School of Science and Technology, Hellenic Open University, Patras, Greece
{panagiotakopoulos,kameas}@eap.gr, borotis@daissy.eap.gr,
fotis.lazarinis@ac.eap.gr
[2] Business School, University of Nicosia, Nicosia, Cyprus
[3] Department of Mathematics, University of Patras, Patras, Greece
kotsiantis@upatras.gr

Abstract. Online learning has developed rapidly in the past decade, leading to increased scientific interest in e-learning environments. Specifically, Massive Open Online Courses (MOOCs) attract a large number of people with respective enrollments meeting an exponential growth during the COVID-19 pandemic. However, only a small number of enrolled learners successfully complete their studies creating an interest in early prediction of dropout. This paper presents the findings of a study conducted during a MOOC for smart city professionals, in which we analyzed demographic and personal information on their own and in tandem with a small set of interaction data between learners and the MOOC, in order to identify factors influencing the decision of starting the MOOC or not. We also applied different models for predicting whether a person previously registered to a MOOC will eventually start it or not, as well as for identifying the most informative attributes for the prediction process. Results show that prediction reached 85% accuracy based only on the number of the first days' logins in the MOOC and few demographic data such as current job role or occupation and number of study hours that the learner estimates he/she can devote on a weekly basis. This information can be exploited by MOOC providers to implement learner engagement strategies in a timely fashion.

Keywords: Dropout prediction · MOOC · Machine learning · Smart cities

1 Introduction

The recent popularity of Massive Open Online Courses (MOOCs), with their huge number of enrolled learners—out of which only a fraction completes their studies successfully [1]—has led to increased interest in dropout prediction. Knowing which learners are likely to quit a MOOC helps distance learning institutions and MOOC providers to

I. Maglogiannis et al. (Eds.): AIAI 2021 Workshops, IFIP AICT 628, pp. 466–475, 2021.
https://doi.org/10.1007/978-3-030-79157-5_38

develop intervention strategies and remedial actions so as to provide personalized support [2]. According to the Coursera 2020 Impact Report [3], more than 80% of the learners report career benefits six months after completing a course on the platform. Given the fact that MOOCs are typically free of charge, people select courses to attend many times without being ready to commit to learning, or even disposing the required study skills. Therefore, MOOCs face increased drop-out rates compared to other learning modalities, an outcome that affects the business models of MOOC providers, their suppliers, and even the learners themselves. In this regard, it is essential for MOOC providers to understand early enough their audiences – ideally from the phase of registration in a course.

Educational Data Mining (EDM) is a proper tool for effectively analyzing learners' behavior and predicting their performance. In this research, we explore the potential of demographic information in tandem with minimal interaction data between learners and MOOCs to identify factors of influence and predict whether a person previously registered in a MOOC will eventually start it or not. Moreover, we examine if such a prediction could be done in a timely manner giving the opportunity to implement early interventions.

The rest of the paper is organized as follows: Sect. 2 reviews recent studies concerning the implementation of machine learning techniques for detecting high-risk students in terms of course completion and performance. Section 3 describes the data collection process, provides a description of the dataset and illustrates the results of data mining and statistical methods. In Sect. 4, we present the classification experiments that were conducted and the results obtained. Finally, the paper concludes by summarizing the main findings of the study and considering some thoughts for future research.

2 Related Work

There are few empirical studies of the relationship between MOOC learners' demographics and completion rates. Guo and Reinecke [4] studied four Open edX MOOCs with data from 140,546 students and found that age was positively correlated with the volume of coverage, which in turn was positively correlated with final grade. On the other hand, Cisel [5] found that employment status had an influence on performance. Unemployed learners achieved higher grades than working students. Morris et al. [6] also found that learners who are older, not working, as well as those with prior online experiences and with prior educational attainment are more likely to complete their course.

Brooks et al. [7] explored the predictive power of learner demographics compared to interaction trace data generated by students in two MOOCs in terms of student success. The authors showed that demographic information offers minimal predictive power in relation to interaction data. Interaction trace data helped models very early in the course. Kizilcec et al. [8] divided students into few stereotypes using a clustering algorithm. The authors used for clustering different data such as students' demographic data, behavior data, geographic and students' course enrollment data. Through their analysis, the authors presented a framework for the conversation about MOOC engagement.

Hone & Said [9] reported a survey study of 379 participants enrolled in a MOOC at university in Cairo. The authors reported that 32.2% out of the participants completed the entire course. The authors stated no significant differences in completion rates by gender and level of study (undergraduate or postgraduate). Qiu et al. [10] proposed a framework to predict learners at risk of not completing the MOOC. This framework used feature selection methods and logistic regression. The presented method searched for the most influential features for training and recurrently tested the predictive power of the model to improve its accuracy. The clickstream data was found as useful. Rizvi et al. [11] compared the influence of demographic characteristics on online learning outcomes using a sample of UK based learners across four Open University online courses. The authors found that region, neighborhood poverty level as well as prior education to have strong influence on overall learning outcomes. However, such influence varied as the course progressed.

The review shows that a number of attempts have been made to predict the completion rates based on demographics, interaction and previous e-learning experiences. The results show that prediction is challenging, especially when relying only on demographics. That is why we employed login data during the first days of the e-course in addition to demographics.

3 Data

Our study was performed in the context of the Erasmus + Sector Skills Alliance project “DevOps: Competences for Smart Cities”[1]. DevOps focuses on equipping current and prospective professionals in municipalities and regional authorities with appropriate competences to support the emerging smart city concept, needs and requirements [12]. Registrations in the DevOps MOOC[2,3] lasted from September 15 to October 15, 2020, while the MOOC started on October 19, 2020. It lasted approximately 3 months and it was structured on a weekly format delivering 1 or 2 training modules (i.e. competences) per week. Each training module – available in English - comprised 2 to 5 learning units, each of which included an automatically graded assessment test. The content delivered to the learners was designed so as to address the European Qualifications Framework level 5, as this is the required level of autonomy and responsibility for smart city professionals. The registration form included a questionnaire asking applicants to provide personal and demographic data notifying them that all data would be acquired and used according to the General Data Protection Regulation Data protection (EU) 2016/679 and the Regulation (EU) No. 2018/1725 to evaluate the quality of the DevOps MOOC. All applicants were requested to provide their consent to store and use this data; otherwise they could skip this questionnaire and proceed with registration only providing their full name and email.

Personal and demographic information included sex, age, nationality, country of residence, mother tongue, education level, current employment status, current job role or occupation, years of experience in the role/occupation, average amount of daily working hours, level of technical English language skills, current digital proficiency, number of under aged children, available amount of study hours on a weekly basis and prior MOOC

[1] https://devops.uth.gr/dev/.

[2] https://devops.uth.gr/dev/about-the-mooc/.

[3] https://smartdevopsmooc.eu/moodle/pages/login.php.

attendance. A total of 961 people applied for the DevOps MOOC, the vast majority of which were from the EU, where the MOOC was promoted. 944 of them provided sociographic information which, after being anonymized, consisted the initial set of attributes along with the number of the first two days logins in the MOOC. Registered learners who did not submit any of the first week's assessment tests (the first training module included 5 assessment tests) are considered as "Not Starting the MOOC" class, while the remaining learners are considered as "Starting the MOOC" class. Figures 1–5 show the impact of various types of demographic and personal information on starting the MOOC or not.

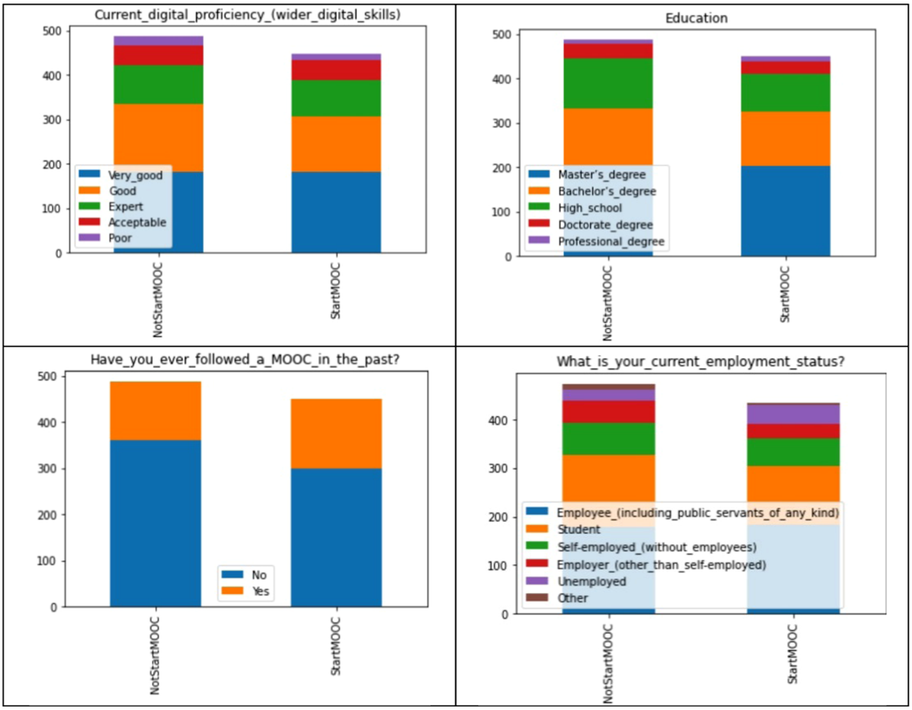

Fig. 1. The impact of some registration information in starting the MOOC.

Figure 1 shows the distribution between those who actually started the MOOC and those who did not start it in relation to their level of education, their employment status and their digital skills and their past e-learning experiences. Approximately 46% started the MOOC. Based on the diagrams of Fig. 1, there are no noticeable differences among the two groups, which corroborates with the findings of the literature review.

Figures 2, 3, 4 and 5 present the data of only those who have not started the MOOC with respect to language competence, digital skills, education level and previous e-learning experiences. Again, it is difficult to draw definite conclusions as the data indicate a balanced distribution among the different groups and among the different skill levels within each group. From Fig. 4, it could be argued that those who have not previously

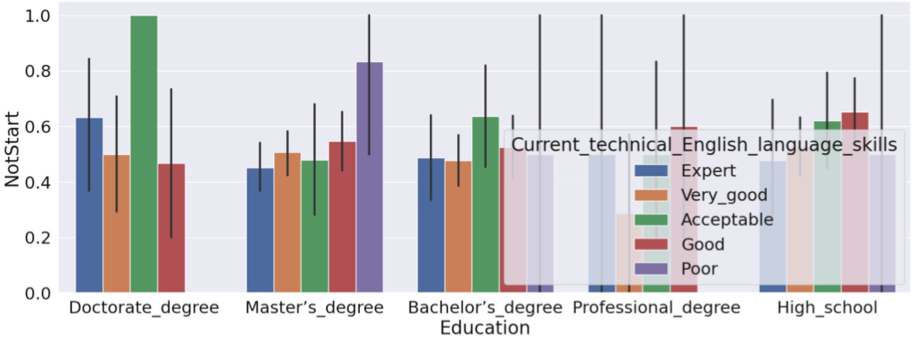

Fig. 2. The impact of current technical English language skills and education level in starting the MOOC

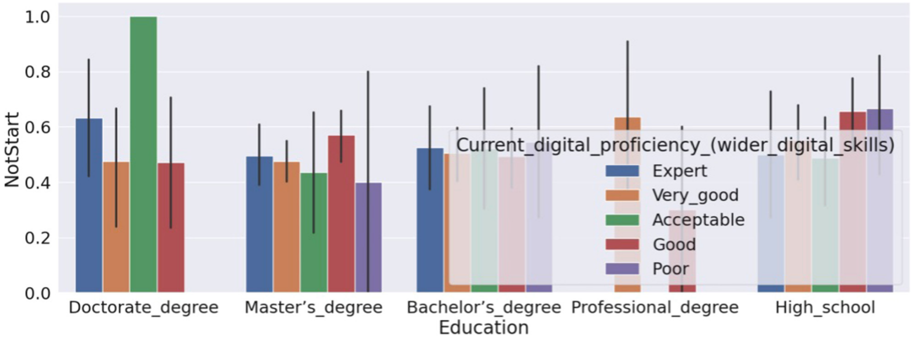

Fig. 3. The impact of current digital skills and education level in starting the MOOC

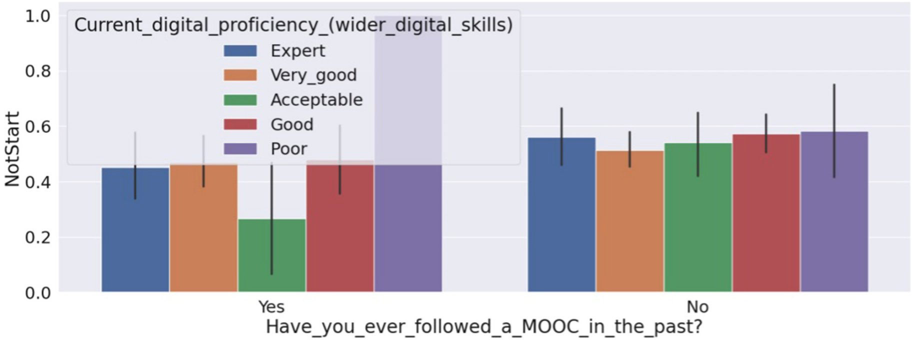

Fig. 4. The impact of current digital skills and the previous experience in MOOCs in starting the MOOC

attended a MOOC have a higher chance of not starting the MOOC. Those with poor digital skills who have already attended a MOOC have a high probability of not starting the MOOC, although this observation needs more thorough analysis and more data to

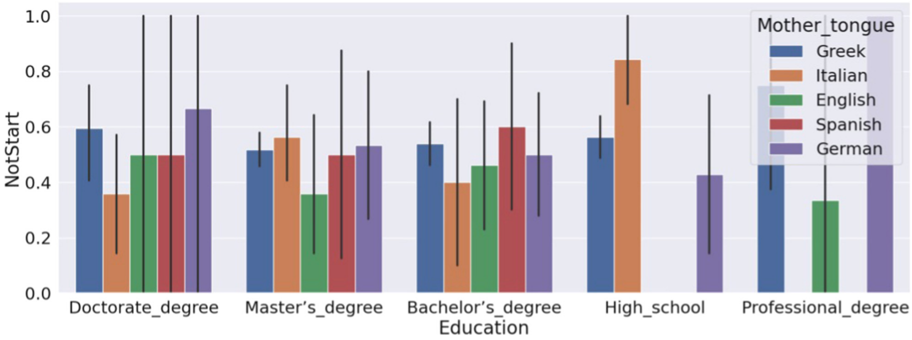

Fig. 5. The impact of mother language and education level in starting the MOOC

verify it. In general, these data should be combined with user logins to understand early in the training period who has a higher probability of not starting the MOOC and to provide more tailored assistance.

An examination of repeated patterns in the learners' data was also performed using association rules [13]. First and foremost, all learners that logged in to the MOOC at least 16 times and declared initially that are able to devote on average at least 5 h per week for their learning, started the MOOC. Second, the learners that logged in to the MOOC less than two times the first (learning) week, did not eventually start the MOOC on a degree of 97%. Third, the learners that logged in to the MOOC less than fifteen times, and their current occupational status was declared as 'employee' (including public servants) or student or unemployed, and could devote between two and twelve hours per week for their learning, started the MOOC. These three rules strengthen the importance of monitoring learners' login activity in the e-course so as to identify whether they shall eventually start their training. Complementary to that, they underline the - even limited - significance of the employment status and the available time for learning as characteristics of the target group that accumulates increased chances to start the MOOC.

4 Prediction Models and Experiments

As we have already said, it is important to recognize the learners that will most probably not start the MOOC at an early stage, in order to implement timely interventions. Thus, the attributes that were collected during registration (i.e. personal and demographic information), as well as those collected during the initial offer period of the MOOC (i.e. first two days logins in the MOOC) were used to identify learners that would potentially not start the MOOC using six different prediction models: Cart decision tree [14], Random forest [15], Multi-layer Perceptron (MLP) classifier [16], Adaboost Classifier [17], Naive Bayes algorithm and Logistic regression [18]. In order to calculate the classifiers' accuracy, stratified ten cross validation was used. Figure 6 depicts the learning curves of the examined learning algorithms.

From the learning curve analysis, we can observe that we can take accurate results using only 100 instances as training set. This rather small number of instances enables MOOC providers with predicting results with even decreased number or resources

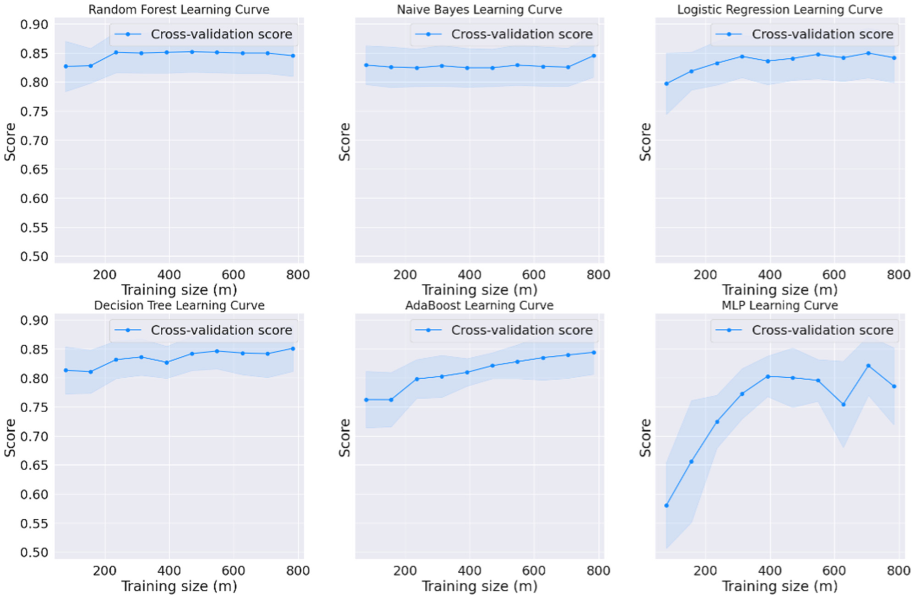

Fig. 6. Classifiers' learning curves

invested for this effort. It must be mentioned that the four examined classifiers produce similar results. There is not one learning algorithm that produces statistically better results than the rest examined methods. However, Random Forest seems to produce slightly better results using less training data.

We then used the Random Forest algorithm for feature importance (see Fig. 7) implemented in scikit-learn [19]. Permutation feature importance is a technique for calculating relative importance scores that is independent of the model used. First, a model is fit on the dataset, such as a model that does not support native feature importance scores. Then the model is used to make predictions on a dataset, although the values of a feature (column) in the dataset are scrambled. This is repeated for each feature in the dataset. Then this whole process is repeated 10 times. The result is a mean importance score for each input feature (and distribution of scores given the repeats), as shown in Fig. 8.

Using ideas from coalitional game theory, the SHAP method [20] computes the Shapley value of a feature, which is the average of the marginal contributions of that feature to all predictions across all permutations of selected features (Fig. 9).

The most informative attributes according to the examined feature importance strategies are user total logins, current job role or occupation and number of hours learners can devote on average per week for learning. The two out of the three features are known before the MOOC starts, enabling MOOC providers to act early enough. Using only the three most informative attributes, the examined learning algorithms lose only about 1% of their accuracy produced with the whole feature space.

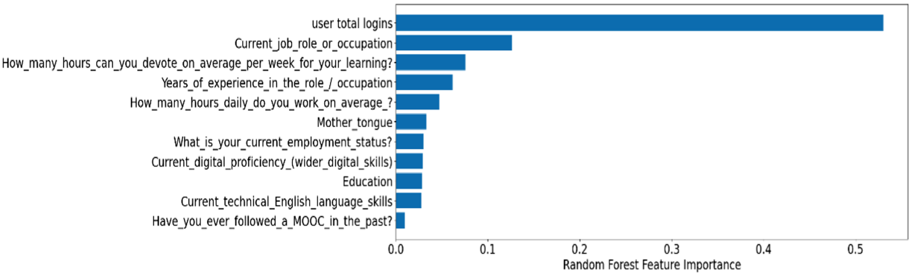

Fig. 7. Random forest feature importance

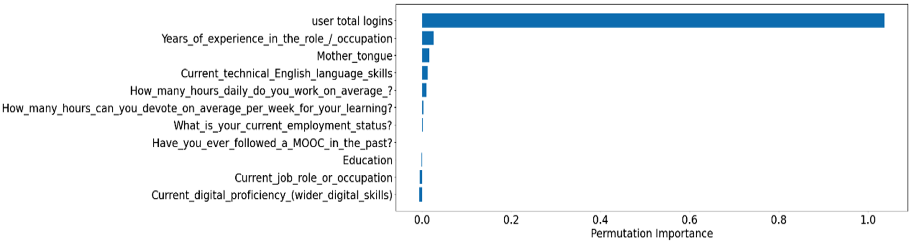

Fig. 8. Permutation feature importance

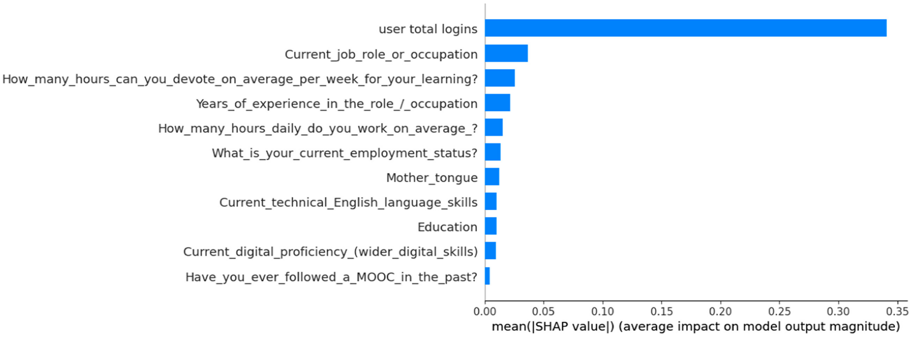

Fig. 9. Mean Shape value

5 Conclusions and Future Work

This paper presented a study concerning the identification of factors affecting the decision of starting the MOOC or not of people who have registered in it. Taking into account learners' logins in the first two days of the MOOC, we employed data mining techniques that revealed interesting patterns linking the actual number of logins on the start of the "MOOC journey" of the learners with characteristics related to their profile. It was shown that the learning algorithms can predict if a learner will start the MOOC after the initial registration. The accuracy reaches 85% based only on the number of the first days' logins in the MOOC and some demographic data such as the current job role or occupation as well as the number of hours that the learner estimates that can devote on average per week for learning.

Being able to know and predict early on the training cycle those more likely to drop out of the course is quite important for MOOC providers to create and implement timely learner engagement strategies. More personalized content and support could be offered, especially for people that relate their learning with career advancement and decreased available time for learning respectively; micro-learning [21] and micro-credentials could be employed for those people as well, opening up education to more people as they support by nature flexibility and inclusiveness. Complementary list of runner up candidates could be exploited, so as to replace those who are likely to drop out; alternative registration policies could be coined; and personalized learning paths could be offered addressing different learning behaviors.

Acknowledgement Supported by the Erasmus+ KA2 under the project DEVOPS, “DevOps competences for Smart Cities” (Project No.: 601015-EPP-1–2018-1-EL-EPPKA2-SSA Erasmus+ Program, KA2: Cooperation for innovation and the exchange of good practices-Sector Skills Alliances).The European Commission’s support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Co-funded by the Erasmus+ Programme of the European Union

References

1. Onah, D.F., Sinclair, J., Boyatt, R.: Dropout rates of massive open online courses: behavioural patterns. In: EDULEARN14 Proceedings, pp. 5825–5834 (2014)
2. Assami, S., Daoudi, N., Ajhoun, R.: Personalization criteria for enhancing learner engagement in MOOC platforms. In: 2018 IEEE Global Engineering Education Conference (EDUCON), pp. 1265–1272. IEEE (2018)
3. Coursera. Coursera 2020 Impact Report. Serving the world through learning. https://about.coursera.org/press/wp-content/uploads/2020/09/Coursera-Impact-Report-2020.pdf. Accessed Mar 2021
4. Guo, P.J., Reinecke, K.: Demographic differences in how students navigate through MOOCS. In: 1st ACM Conference on Learning@scale, pp. 21–30. ACM (2014)
5. Cisel, M.: Analysing completion rates in the First French xMOOC. In: EMOOCs 2014: European MOOCs Stakeholders Summit. Proceedings. Research Track, pp. 26–32 (2014)
6. Morris, N.P., Swinnerton, B.J.D., Hotchkiss, S.: Can demographic information predict MOOC learner outcomes? In: Experience Track: Proceedings of the European MOOC Stakeholder. eMOOCs Conference, Mons, Belgium (2015)
7. Brooks, C., Thompson, C., Teasley, S.: Who you are or what you do: comparing the predictive power of demographics vs. activity patterns in massive open online courses (MOOCs). In: 2nd ACM Conference on Learning@ Scale, pp. 245–248. ACM (2015)
8. Kizilcec, R.F., Piech, C., Schneider, E.: Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 170–179. ACM (2013)
9. Hone, K.S., El. Said, G.R.: Exploring the factors affecting MOOC retention: a survey study. Comput. Educ. **98**, 157–168 (2016)
10. Qiu, L., Liu, Y., Liu, Y.: An integrated framework with feature selection for dropout prediction in massive open online courses. IEEE Access **6**, 71474–71484 (2018)

11. Rizvi, S., Rienties, B., Khoja, S.A.: The role of demographics in online learning; a decision tree based approach. Comput. Educ. **137**, 32–47 (2019)
12. Iatrellis, O., Panagiotakopoulos, T., Gerogiannis, V.C., Fitsilis, P., Kameas, A.: Cloud computing and semantic web technologies for ubiquitous management of smart cities-related competences. Educ. Inf. Technol. **26**(2), 2143–2164 (2020). https://doi.org/10.1007/s10639-020-10351-9
13. Abdullah, Z., Herawan, T., Ahmad, N., Deris, M.M.: Mining significant association rules from educational data using critical relative support approach. Procedia-Soc. Behav. Sci. **28**, 97–101 (2011)
14. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth, Belmont (1984)
15. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
16. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986)
17. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
18. Ng, A.Y., Jordan, M.I.: On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. Adv. Neural Inf. Process. Syst. **2**, 841–848 (2002)
19. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
20. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874 (2017)
21. Emerson, L.C., Berge, Z.L.: Microlearning: Knowledge management applications and competency-based training in the workplace. Knowl. Manage. E-Learn. **10**(2), 125–132 (2018)