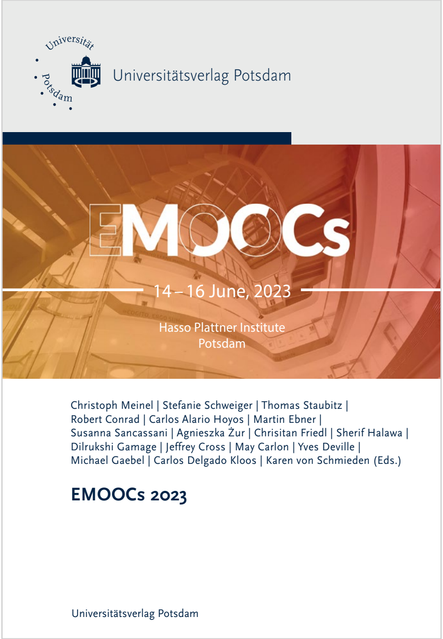


Universität Potsdam
Universitätsverlag Potsdam

EMOOCs

14 – 16 June, 2023

Hasso Plattner Institute
Potsdam

Christoph Meinel | Stefanie Schweiger | Thomas Staubitz |
Robert Conrad | Carlos Alario Hoyos | Martin Ebner |
Susanna Sancassani | Agnieszka Żur | Chrisitan Friedl | Sherif Halawa |
Dilrukshi Gamage | Jeffrey Cross | May Carlon | Yves Deville |
Michael Gaebel | Carlos Delgado Kloos | Karen von Schmieden (Eds.)

EMOOCs 2023

Universitätsverlag Potsdam

Suggested citation:
Mehdi Khaneboubi: Visualizing students flows to monitor persistence, In: Christoph Meinel, Stefanie Schweiger, Thomas Staubitz, Robert Conrad, Carlos Alario Hoyos, Martin Ebner, Susanna Sancassani, Agnieszka Żur, Christian Friedl, Sherif Halawa, Dilrukshi Gamage, Jeffrey Cross, May Kristine Jonson Carlon, Yves Deville, Michael Gaebel, Carlos Delgado Kloos, Karen von Schmieden (Eds.): EMOOCs 2023 : Post-Covid Prospects for Massive Open Online Courses - Boost or Backlash?, Potsdam, Universitätsverlag Potsdam, 2023, S. 121–131.
DOI https://doi.org/10.25932/publishup-62390

Visualizing students flows to monitor persistence

Mehdi Khaneboubi

Université Paris Cité, EDA, F-75006 Paris, France
mehdi.khaneboubi@gmail.com

Founded in 2013, OpenClassrooms is a French online learning company that offers both paid courses and free MOOCs on a wide range of topics, including computer science and education. In 2021, in partnership with the EDA research unit, OpenClassrooms shared a database to solve the problem of how to increase persistence in their paid courses, which consist of a series of MOOCs and human mentoring. Our statistical analysis aims to identify reasons for dropouts that are due to the course design rather than demographic predictors or external factors. We aim to identify at-risk students, i.e. those who are on the verge of dropping out at a specific moment. To achieve this, we use learning analytics to characterize student behavior. We conducted data analysis on a sample of data related to the "Web Designers" and "Instructional Design" courses. By visualizing the student flow and constructing speed and acceleration predictors, we can identify which parts of the course need to be calibrated and when particular attention should be paid to these at-risk students.

1 Introduction

Founded in 2013, OpenClassrooms is a French online learning company that provides both paid courses and free MOOCs covering a wide range of topics, including computer science and education. In 2021, OpenClassrooms partnered with EDA research unit to share a database, aiming to address the challenge of studying persistence in their paid courses, which consist of a series of MOOCs, projects, and human mentoring. Accurately predicting dropouts early allows course designers and educators to adjust their teaching methods. This raises the question of finding predictors that can explain dropout and persistence.

Kizilcec et al. (2013) [8] presented a list of dropout predictors based on student activity features to predict which students are at risk of dropping out in three computer science MOOCs. They identified four engagement profiles based on variables such as demographics, forum participation, video access, and overall experience reports. Halawa and al. (2014) [7] demonstrated that performance on assessments, skipping assessments, skipping videos, and falling behind on watching video lectures are predictors of the likelihood of dropping out. Wei et al. (2023) [10] explained perceived learning outcomes with several predictors, including course design, interaction with instructors and peers, engagement in learning activities, and application of cognitive and metacognitive learning strategies.

Numerous researchers have described machine learning methodologies that predict dropout in online training [9, 12, 5, 3]. They define at-risk students according to the needs of machine learning methods. They center the definition of variables to be explained in relation to the availability of descriptors in their data and the requirements of machine learning algorithms. Evans and Baker (2016) [6] assess several definitions of persistence based on data collected from ten Coursera MOOCs. They observed different values for persistence according to their definition, but the pattern of high variance across measures remained the same. How persistence rates are calculated reflects differences in student goals for participating in a MOOC. They concluded that important factors explaining persistence are: first, the time when the video are published and the lexicon used to describe it, and then, the level of the course: as in university education, students tend to avoid assignments that require high levels of writing.

In a preprint document [4], Chibaya et al. (2022) present categories of situations for unsuccessful higher education students, which include:

- Permanently quitting from studies (dropout)
- Temporarily discontinuing studies with the hope of re-registering (stop-out)
- Responding to chronic stress through emotional and physical exhaustion (burnout)
- Enduring through a study program without success to the exam (failing)

They clearly present the situations that could occur but do not define the "at-risk" concept. In fact, being "at-risk for a student" depends on the student's profile in relation to the content being taught and the pedagogical setup of the training. It seems difficult to define it in an universal way, but it could be done by describing the training and the profiles of the students.

However, we can search for the specific behaviors of students who are highly likely to drop out. What are the first signs of weakness? Atif et al. (2020) investigated the perspectives of teachers regarding an early alert system in face-to-face training using Moodle [1]. The teachers' statements revealed that there is no standard approach to identifying at-risk students, including class attendance, assessment submissions, assessment types, and forum participation. However, the most commonly reported predictors by teachers are the submission of first assessments and assignments, as well as the achievement of certain grades. Wolff et al. (2013) [11] made a data analysis on click behavior of students in online training. Their primary finding is that a change in student behavior is the best predictor of dropout. If a student stops logging into the platform, there is a higher likelihood of dropout than for a student who initially has a low level of activity.

The context of our study differs from the researches cited above. Firstly, the OpenClassrooms training is entirely online and is based on a series of MOOCs that are not compulsory, but rather available as resources. Secondly, learners work individually and have flexible time frames to start and complete their training. Unlike traditional distance learning courses, students are not grouped into cohorts. As a result, we adopt Bruillard's (2017) [2] perspective that MOOCs should be considered as learning materials, rather than as online courses. In our context, access, attendance, and recognition have a different nature compared to traditional distance learning courses. In essence, MOOCs are dynamic websites made of text, videos, audios, social interactions, online communities, video conferences, and other features. This viewpoint of MOOCs enables us to differentiate the factors and standards that constitute the training in a novel manner.

2 Data and Method

The OpenClassrooms training program is designed to allow students to start and progress through the training at their own pace. To successfully complete the program, students must complete seven projects[1], which are evaluated by a mentor and defended in front of a jury. These projects must be completed in sequential order. Mentors are providing guidance and are assessing the quality of the work before the student is permitted to defend his work. Mentors are experts in their respective fields, possessing a deep understanding of the course content. Among the courses offered, the *Web Developer* course is the most popular, while the *Pedagogical Engineer* course has a smaller audience.

A unique database was accessed for the *Web Developer* training, which provided a sample dataset of 2995 students described by 59 variables. The primary explanatory variable in our analysis is the status variable, as shown in Figure 1.

There are three categories that describe dropout: churners, expired, and abandoned. Initially, we used these categories to define and identify at-risk students, i.e. those for whom the probability of dropping out could be predicted. Subsequently, we explored statistical associations between these students and other variables with

[1] The projects are: "Take charge of your web developer training", "Turn a template into a web site", "Make a web page dynamic with CSS animations", "Optimize an existing website", "Build an e-commerce website", "Build a secure API for a restaurant review application", "Create an corporate social network".

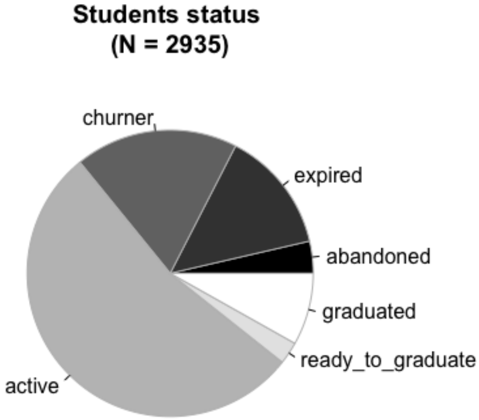


Figure 1: Student status for the *Web Developer* training

the ultimate objective of establishing predictors of dropout based on their online behavior.

Defining and identifying at-risk students is crucial for characterizing the problem and designing indicators to inform educational action. Exploring statistical associations between these students and variables in the database offers the possibility of constructing statistical predictors.

It is relevant and important to view the dataset as a snapshot of the training dynamics. The training process can be conceptualized as a progression of students navigating through projects, resources, and assignments. Therefore, it is necessary to identify and characterize typical movements of students, regular patterns in the course of training, and similar behavior that reveal organization and the opportunity for intervention at a specific moment. The objective of this study is to identify at-risk students and make possible special support when they require it the most.

3 Results

3.1 Representing at-risk students

Crossing the variable that characterizes the students' situation (active, certified or almost, and dropped out) with the start date of the training permit to draw Figure 2 diagram.

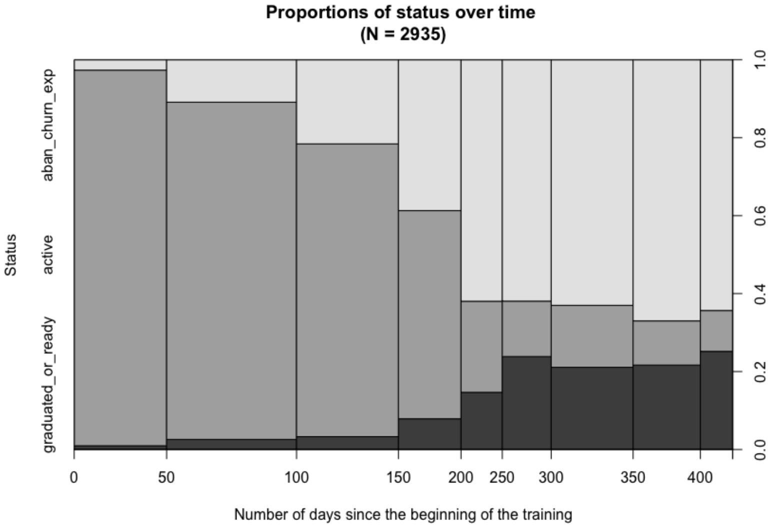
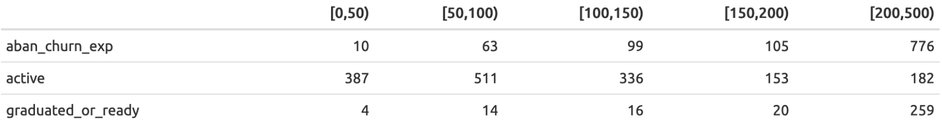


Figure 2: Proportion of students according to their status over time

The lighter gray area in Figure 2 represents the proportion of students who have dropped out, while the darker gray area represents those who have been certified or are in the process of certification. The middle area illustrates active students. The diagram shows that the proportion of students dropping out increases from 0 to 200 days after the start of the training, and thereafter stabilizes. When viewed as a snapshot in a continuous flow of collective movements, "at-risk" students are those represented just above the lower part of the light gray area. For these students, the probability of dropping out in the next period is higher than for others.

Table 1 corresponds to Figure 2 and shows that 267 students dropped out between 50 and 200 days into the training.

Table 1: Student status according to time since the beginning of the training



	[0,50)	[50,100)	[100,150)	[150,200)	[200,500)
aban_churn_exp	10	63	99	105	776
active	387	511	336	153	182
graduated_or_ready	4	14	16	20	259

This data can be interpreted as temporal milestones: the dropout rate rises every 50 days of training and plateaus at 200 days. These time intervals can be viewed as the most critical points for intervention to prevent dropout.

3.2 Visualizing pathway

Figure 3, presented below, displays, in row, projects that must be completed in order to succeed in the course. Time periods are represented in columns (0 to 50 days since the start, 50 to 100 days, etc.). Student status is indicated by colored dots: active students are represented in green, while those who have been inactive in the course are colored in red.

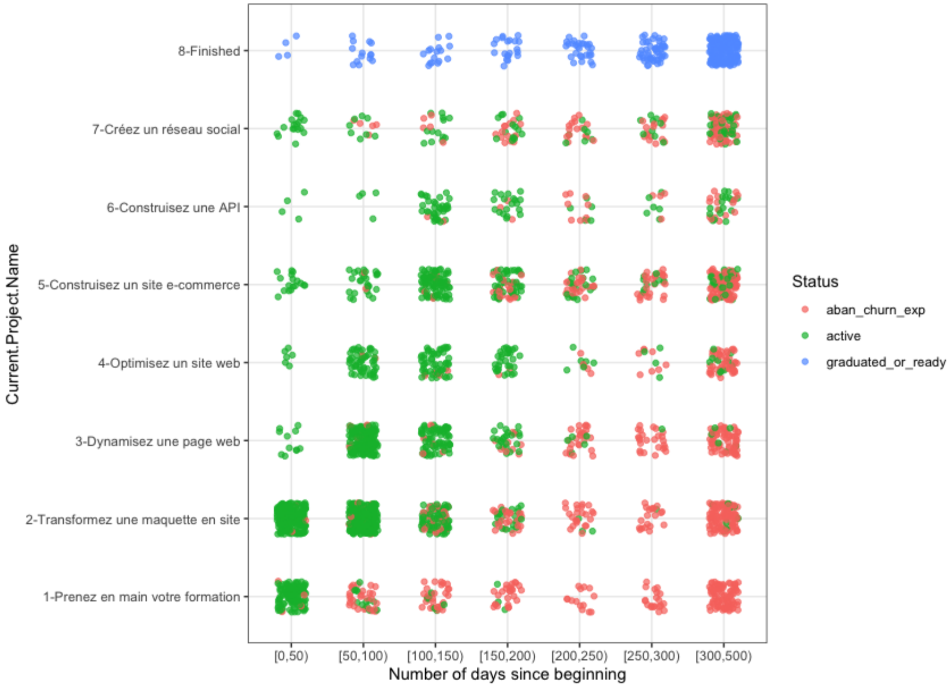


Figure 3: Distribution of students according to their achievement, the number of days in the training, and their status.

This scatter plot has to be seen as a checkerboard-style board. The bottom left square represents students who started their training less than 50 days ago and must complete the first project named “1-Taking in charge of your training”. It is mostly green, indicating that the students are active. From a flow or stream perspective, this box is the source of the flow, and the flow direction is towards the highest row in which all projects have been completed, represented in blue. As time progresses, each dot will move to one of the three consecutive squares: on the upwards or top right square if the first project is made, on the right square if not.

In this perspective, the overall flow of students starts from the southwestern square and proceeds northward with an attraction towards the east of the representation. The more vertical a student’s pathway is, the faster they will have completed the projects. The more horizontal a student’s path is, the longer their training will take, and the more likely they are to drop out. This is evident by observing the strong concentrations of green dots that remain in the first three columns, as well as the strong concentrations of red dots located in the southeastern part of the graph, in the lower part of the last three columns.

Therefore, we can define at-risk students as those whose trajectory is more oriented towards the east of the graph, while successful students are those whose trajectory is more oriented towards the north.

3.3 Instantaneous and average speeds

We have re-established our collaboration with OpenClassrooms to improve the characterization of student flows. We obtained data from the *Educational Engineer* course and realized that the variable used to define student activity in our previous analysis was not very robust. Defining online or offline activity for infrastructure services is not straightforward. Therefore, we chose to define activity ourselves using a new database that is updated every three weeks. This allows us to extract training data and generate reports every three weeks. Using these databases, we calculated the instantaneous speed per student, as shown in Figure 4 below, which should be viewed as a checkerboard similar to Figure 3.

In this figure, green dots represent students with the highest speed, orange dots indicate the slowest, and grey dots show students without a speed calculation so far. We observe a pattern similar to a coastline from the southwest to the northeast of the graph. The fastest students in the last month is colored in green and is the closest part of this coastline, while the slowest students are located in the southeast part, colored in orange.

From a methodological standpoint, this calculation is possible with at least two datasets extracted at different moments during the training. With at least three extractions, we can calculate an average speed, as shown in Figure 5.

In Figure 5 purple dots represent at-risk students. They are progressing in their training, but at a slower pace than the orange and green students. Their average

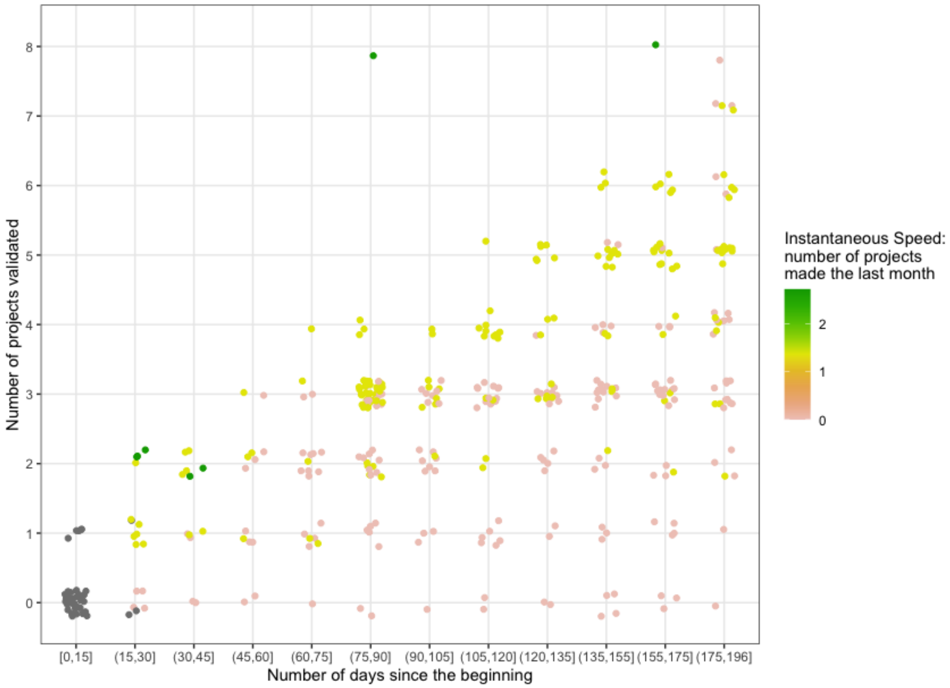


Figure 4: Distribution of students according to their achievement, the number of days in the training, and their instantaneous speeds

speed is approximately half a project per month. The flow of these students is mainly directed towards the east, while the orange and green flows are directed towards both the north and the east.

By using three datasets, we can calculate the instantaneous acceleration of the students. With a fourth dataset, we can calculate both the average acceleration and an indicator of regularity or consistency to estimate the reliability of the speed. This will enable us to determine whether a student's speed and acceleration remain consistent across all extractions.

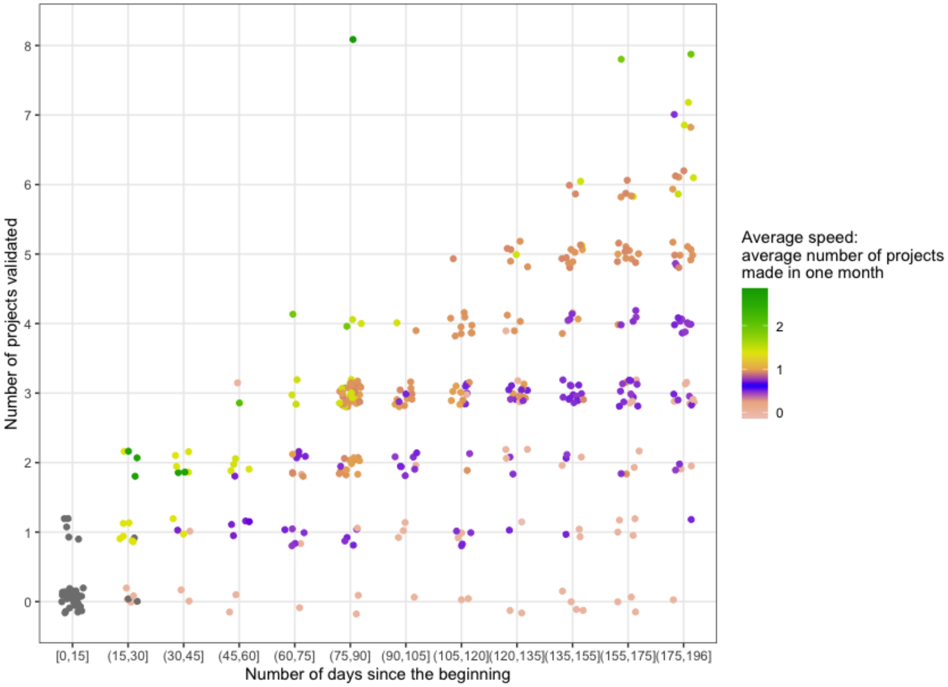


Figure 5: Distribution of students according to their achievement, the number of days in the training, and their average speed

4 Conclusion

This study is still in its early stages and is ongoing. It shows that by using two variables (starting time and number of projects completed) from at least two datasets extracted at different moments during training, we can visualize the streams of students. By calculating instantaneous and average speed, we can identify students who need special or extra attention from course educators.

Our visualization of student streams and course dynamics permits the characterization of student activity without demographic or learning predictors. By not using those predictors to define at-risk students, we can use them as explanatory variables. This method avoids the issues of defining dropout, as shown by [7].

Calculating the speed of students enables the representation of streams. If the results from [11] apply to our context, we can predict the dropout of at-risk students by identifying a deceleration in the students' training.

The checkerboard graph can be summarized with a tree, which will provide an easy way to create typologies and construct profiles.

Our next step is to calculate acceleration and regularity indicators based on four extractions. These indicators will show other aspects of the streams, such as whether borderline students are regular or if their pathways are irregular. Then, we will seek correlations with logistic regression. Finally, we will attempt to predict the behavior of at-risk students using the Random Forest and Gradient Boosting algorithms.

Two predictors will be used to do so: factors that are the responsibility of the training designers and those that are the responsibility of the learners. Once this distinction is established, we will be in a position to answer common research questions on dropout by identifying predictors based on students' navigation through the course concerning its design and the potential to act on them.

References

[1] A. Atif, D. Richards, D. Liu, and A. Bilgin. "Perceived benefits and barriers of a prototype early alert system to detect engagement and support 'at-risk' students: The teacher perspective". In: *Computers & Education* 156 (2020), page 103954. DOI: 10.1016/j.compedu.2020.103954.

[2] É. Bruillard. "Mooc une forme contemporaine de livres éducatifs. De nouveaux genres à explorer ? Distances et médiations des savoirs". In: *Distance and Mediation of Knowledge* (2017).

[3] J. Chen, B. Fang, H. Zhang, and X. Xue. "A systematic review for MOOC dropout prediction from the perspective of machine learning". In: *Interactive Learning Environments* 0 (2022), pages 1–14. DOI: 10.1080/10494820.2022.2124425.

[4] C. Chibaya, A. Whata, K. Madzima, G. Rudolph, S. Verkijika, L. Makhoere, and M. Mosia. "A scoping review of the 'at-risk' student literature in higher education". In: (2022). DOI: 10.1101/2022.07.06.499019.

[5] K. Chui, D. Fung, M. Lytras, and T. Lam. "Predicting at-risk university students in a virtual learning environment via a machine learning algorithm". In: *Computers in Human Behavior* 107 (2020), page 105584. DOI: 10.1016/j.chb.2018.06.032.

[6] B. Evans and R. Baker. "MOOCs and Persistence: Definitions and Predictors". In: *New Directions for Institutional Research* 2015 (), pages 69–85.

[7] S. Halawa, D. Greene, and J. Mitchell. *Dropout Prediction in MOOCs using Learner Activity Features*. 2014.

[8] R. Kizilcec, C. Piech, and E. Schneider. “Deconstructing disengagement: analyzing learner subpopulations in massive open online courses”. In: *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. 2013, pages 170–179. DOI: 10.1145/2460296.2460330.

[9] J. Kuzilek, M. Hlosta, D. Herrmannova, Z. Zdrahal, J. Vaclavek, and A. Wolff. “OU Analyse: analysing at-risk students at The Open University”. In: *Learning Analytics Review* LAK15-1 (2015), pages 1–16.

[10] X. Wei, N. Saab, and W. Admiraal. “Do learners share the same perceived learning outcomes in MOOCs? Identifying the role of motivation, perceived learning support, learning engagement, and self-regulated learning strategies”. In: *The Internet and Higher Education* 56 (2023), page 100880. DOI: 10.1016/j.iheduc.2022.100880.

[11] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek. “Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment”. In: *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. 2013, pages 145–149. DOI: 10.1145/2460296.2460324.

[12] L. Zhang and H. Rangwala. “Early Identification of At-Risk Students Using Iterative Logistic Regression”. In: *Artificial Intelligence in Education, Lecture Notes in Computer Science*. 2018, pages 613–626. DOI: 10.1007/978-3-319-93843-1_45.