IEEE CITS 2018

2018 International Conference on Computer, Information and Telecommunication Systems

July 11-13, 2018, Colmar, France

Editors:
Mohammad S. Obaidat, Fellow of IEEE and Fellow of SCS, Pascal Lorenz, Kuei-Fang Hsiao
Petros Nicopolitidis, Daniel Cascado-Caballero

Technical Sponsors:

IEEE

IEEE ComSoc™ IEEE Communications Society

SCS

UNIVERSITÉ HAUTE-ALSACE

For this compilation:
ISBN: 978-1-5386-4075-3

Analysis of a short on-line course through logged data recording by a self-developed logging module

Péter Esztelecki
Faculty of Science and Informatics
University of Szeged
Szeged
epeter@inf.u-szeged.hu

Gabor Korosi
Faculty of Science and Informatics
University of Szeged
Szeged
korosig@inf.u-szeged.hu

***Abstract*—Online education has gained a wide popularity in today's global information boom. Prominent universities offer more and more online courses with modern audio-visual content, which have become available for almost everyone. The courses can be completed self-paced allowing for much more flexibility. Such a learning approach has already reformed higher education and seeks ways to penetrate into the realm of the secondary and primary education. Our team conducted a research in the higher classes of different primary schools in the Province of Vojvodina, Serbia. Participants of the Hungarian minority took part in a course named „Conscious and Safe Internet Use" and obtained a valuable knowledge with the help of videos and optional course materials. Student activities were all recorded with the help of a special self-developed logging software for later processing. Data were processed by statistical and data mining methods. According to the values of correlation coefficients and R Square Statistical value it can be stated, that those participants who scored well at the pre-test stage expectedly had better results during the final testing stage. The number of video views, the age of students and the size of their hometown had also influenced the outcome of the tests. The detailed findings are presented in this paper under the Results chapter.**

Keywords—E-learning, Educational Data Logging, Log Data Analysis

I. INTRODUCTION

E-learning education supports computer learning and allows for a self-paced timing, which means participants do not need to wait for the beginning of school lessons but can acquire new knowledge self-paced in their free time. According to I. Elaine Allen and Jeff Seaman, those courses are identified as online courses whose content amount of at least 80 percent is delivered online [1]. Online courses allow for a cheaper and widely accessible education since broadband internet connection and infrastructure are given in many countries [2]. It is thus a fertile ground for the spreading of online courses.

Initial online courses contained only textual data with supporting charts and tables since video content could not be incorporated due to slow internet connection. Though, we must mention that in the field of distant learning, there were some trials with educational TV programs in the 60's. Nowadays, the majority of online courses contain audio-visual data with less textual content which aim at targeting learning through seeing and hearing. [3]. To gain a deeper overview of the entire online course and its completion success, it is not enough to observe the structure of the course material, content, acquired credits, or the time spent in on the platform but more importantly we have to shed some light on the participants` overall activities on the site. There are a couple of programs developed to suit this task to log student activities. In recent years, a relatively large body of research has been dedicated to characterizing students' behaviour on the basis of their logged activities [4]. Since, logging student activities require the storage of a huge amount of data we can safely talk about them as a form of Big Data which require statistical analysis and Data Mining approach. Big Data analysis is in close connection with data mining which connects the field of data mining, statistics, and artificial intelligence; furthermore, it aims at touching upon hidden relationships that would help gain new knowledge not only in the commercial field but also in the area of economics providing useful information. Danah Boyd and Kate Crawford created the following definition of Big Data: "It is the kind of data that encourages the practice of apophenia: seeing patterns where none actually exist, simply because massive quantities of data can offer connections that radiate in all directions" [5]. The concept of Educational Data Mining is well defined by Cristóbal Romero and Sebastion Ventura: it is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context. EDM uses computational approaches to analyze educational data in order to study educational questions. The EDM process converts raw data coming from educational systems into useful information that could potentially have a great impact on educational research and practice. [6].

In this paper, we wish to present a self-developed MOOC program which saves student activities. Out of this volume, our aim is to demonstrate some information and useful findings that we came upon with after a thorough study. With the rise in number of online education platforms and the development of MOOC's, all the data gathered this way assign a completely new meaning to the course design. Big data allow for very exciting changes in the educational field that will revolutionize the way students learn and teachers teach [7].

II. BACKGROUND

According to Hershkovitz and Nachmias when we analyse logged data we tend to refer to three main parameters: the action taken, who took it and when [4]. Upon building our

978-1-5386-4599-4/18/$31.00 ©2018 IEEE

system we also took these three parameters into account. The identification of a participant was an easy task, since all the learning material were only available after logging into the platform, thus we could link an identification number to a student. The time stamp was always at disposal from the server time making time logging also an important condition for a precise analysis. To record the action taken by a participant was a more difficult task, since it requires JavaScript event handlers.

Regarding courses the drop-out rate also plays a significant role. Carr states that it is widely agreed that drop out in online learning is higher, often by 10–20 percentage points, than in traditional learning [8]. Not only these value are important to be defined but also it is crucial to answer the WHY questions which require the study and analysis of numerous data [9]. Some of the following factors have impact on the drop-out rate: the structure of the course, the structure of the videos of a lesson (length, quality, etc.), the personality of the lecturer, socio-demographical issues, etc. By taking a look at the logged data with Data Mining methods, we can gain insight into new pieces of information.

III. Methodology

A. Population

After managing to put the final version of the logging server together in May, 2017, we announced a course with the title: Conscious and Safe Internet Use. Primary school participants from classes 5,6,7, and 8 took part in the course with a Hungarian minority background in the northern Province of Vojvodina in Serbia. Altogether 1370 learners logged into the system when the course was live. Finally, we were left with 1076 participants since some who initially signed up to the course did not fill in every test or just logged in once during the whole course. 54% of the students were females. The birth dates of the learners vary according to the following dates: 2001: 1,02%, 2002: 15,24%, 2003: 30,67%, 2004: 26,95%, 2005: 23,23 , 2006: 2,88%.

B. Curriculum and Tests

The course named Conscious and Safe Internet Use began with a pre-entry test to assess students' knowledge, which were filled in under teacher supervision in the informatics lab of a school where the participants belonged. Learners were able to provide their gender, date of birth, name and place of their school and had to answer 10 questions in connection with the upcoming course. These questions were present in the end term tests, as well, which allowed to check how much they improved by the end of the course. The learning material consisted of three main parts: Digital Footprint, Safe Use of Internet, and Online Mobbing. Each module contained a video lesson (the professional videos were shot with the help of the green box technic and were 13-14 minutes long) and there were several references to relevant literature. The time frame to watch a video and to learn the material was one week. Students thus could use their own time management to watch a video and to prepare for the end term test. Finally, they had to fill in three tests connecting to three curriculum modules, each containing 10 questions.

C. The Logging Module

With the Software Engineering Department of the University of Szeged, we developed a Moodle based logging module, which is capable of recording all the activities taken in the system for further analysis. The client side JavaScript software recognizes student activities (for example, clicking, mouse movement, video play) and stores them in an event buffer to optimize data traffic and then they are sent and saved on the server in a Mongo Database (see Fig. 1.).

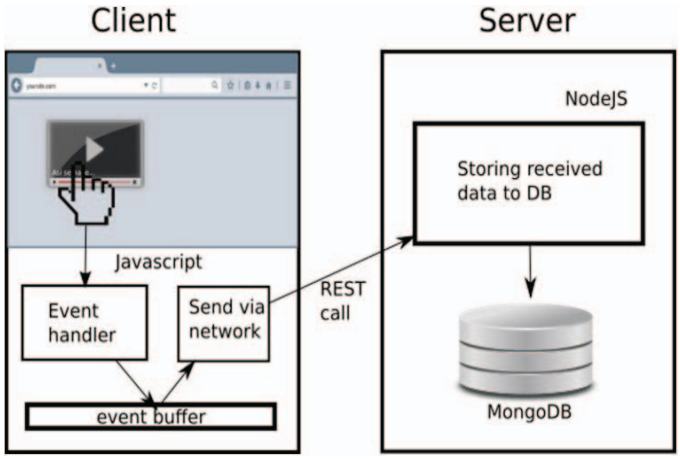

Fig. 1. Event logging

The system creates a record of the event's timestamp in every single case. It saves on which site the event occurred and which user triggered the event (before logging into the Moodle system, the users are identified as Guest users). The module records the coordinates of the mouse movement from to which the user moved the cursor, which pixel was clicked, from which position to which he scrolled, the opening and closing of a page, when was the page put into focus or was blurred, the click in an input field and characters typed in (in case of a password the characters are always hidden). Furthermore, the module records a downloaded file, a started or stopped video, a resized screen, audio volume change, and video seek. The script in a JSON file (JavaScript Object Notation) generated by the Mongo database regarding mouse movement is the following:

{ "type" : "videoPlay",
"data" : { "actualTime" : 14.878231, "videoId" : "video1", "totalTime" : 844.881269, "src" : "https://elearning.szte.hu/elearningdata/videok/tudatos_es_biztonsagos_internethasznalat/1-A_digitalis_labnyom.mp4" },
"time" : "2017.05.25. 12:29:29.855",
"page": "https://tanul.sed.hu/mod/szte/course.php?id=3...,
"user" : 2365,
"_id" : { "$oid" : "5926b20bf52e3962c1f8587c" } }

The first parameter describes the event type then come data about the actual time, id, total time and source of the video, following the timestamp, page address and the user id, and finally an entry identification code which is generated by the Mongo database system.

D. The Log File

All the events during the course, namely the 2.425.484 interpretable lines were recorded into a 674,3MB long JSON file. Most of the data processing software available is incompatible with the JSON format, while the CSV (Comma Separated Values) file types can be easily handled. Thus, we designed a PHP based software which converts logged data (see Figure 2) and the output is saved in the following way into the user-1548.csv file:

mousemove,310,1296,31,18,38.98,2017.05.29. 08:06:04.575,https://tanul.sed.hu/mod/szte/course.php?id=3...

(Remark: the last parameter of the file containing all the user information is the user id). The index page of the program allows for the choice of events to be saved into the output file using check boxes; furthermore, it is possible to choose one CSV file for every user or individual files for all the users, or the compression of the consecutive mouse movement. After hitting the Generate! button, the program reads the input file by lines and puts the entries into the output file.

Fig. 2. Screenshot of the software for conversion

After running the code, the CSV files are closed and a summary file is also created with the number of processed data lines, events taken into account, the amount of time needed for conversion, the size of the input file and the output file respectively. Furthermore, the program attaches and saves events that belong to an individual user. If the mouse move compression was also ticked, the program creates a CSV file for every user where all the consecutive mouse movements are saved in the following format:

mousemovecompr,5,281,1278,308,1300,85.24,2017.05.29. 08:06:04.075,2017.05.29. 08:06:06.075, https://tanul.sed.hu/mod/szte/course.php?id=3...

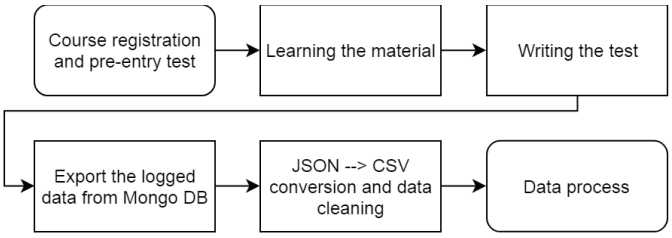

Fig. 3. Flowchart of the process

The compression of the saved data is important because 75% of all the recorded data are mouse movements, which can be easily contracted, thus 5 times less mouse move event is saved into the output file allowing for a faster processing.

Participants had to provide some personal initial data: gender, date of birth, place of residence and the name of the school. Some false data were also recorded into the database but they were easily corrected by looking at peer data of those who filled in the paper in the same time. Regarding town names, we had to manually control and correct typical mistakes or incompletion, for example learners mistyped or were not able to write down correctly the name of their hometown, however, the most typical mistake was providing the Serbian alternative of the town names. Subsequently, we concluded that a drop-down menu would have eliminated this problem, since out of the 31 town names, the participants came up with about 100 false entries. The number of citizens has been added later to town names based on the census from 2011 and divided into four categories.

The supervising teachers had also an access to the system allowing them to watch the videos. The system, additionally, recorded the events created by the administrators. Later, these records were removed as they would lead to false conclusions. Actions taken by the Guest users were also deleted as they were rated as uninterpretable. The result files generated by the Moodle had to be also reprocessed: the points were summed up and instead of showing every answer a,b,c, or d options were recorded. Time spent on a test was not saved as a string with n minutes and m seconds but was converted into an integer with only the sum of seconds. After cleaning and converting the database, we joined the tables with SQL commands to suit further statistical and Data Mining processing. By joining the tables, we were able to assign events to individual users and to find test related data or to search for personal learner information.

IV. Results

The conversion and processing of the log file were followed by statistical and Data Mining processing.

Out of 1076participants, 632 learners filled in all the 4 tests (i.e. one pre-test and three final tests). Log data clearly show that those participants who did not fill in every test were less active during the course (see Table 1.).

TABLE I. STUDENT ACTIVITY (N=1076)

	mousemove	bflu	click	scroll	vidplay	vidseek
4 tests completed (n=632)	1560.96	21.29	55.61	281.22	2.49	0.96
<4 tests completed (n=444)	870.6	14.82	31.02	151.69	1	0.38

The 632 learners who completed the pre-test and the three final tests had twice as more mouse movement, blur, focus, load, and unload (bflu) events. The same results were achieved with click and scroll events, while there is even a higher gap in the frequency of video play and video seek options.

Female participants had better scores at every test. The difference is in average 0.57 point (see Table 2.). The t0avg column shows the average pre-test and the t1-t2-t3avg indicates the final three tests.

TABLE II. TEST RESULTS (N=1076)

	t0avg	t1avg	t2avg	t3avg
Male (n=583)	5.05	5.03	5.44	5.12
Feale (n=493)	5.65	5.25	6.2	5.83

To pass the course, the participants had to score at least 5 points. The pre-test did not have such a prerequisite since only the level of knowledge was assessed. The course was successfully completed by 282 learners (175 girls, 107 boys) whose video play scores were 3.67 while those users who did not reach the minimum 5 points, the value was only 1.55.

By studying the correlation factors, the relationship is even more significant (at the level 0.01) when the number of clicking and mouse movement (0.475), clicking and video play (0.217), mouse movement and video play (0.317) are compared. With the analysis of the correlation factor of the number of clicking and mouse movement, their value remain under 0.1. The correlation factor between video play and the tests remain also significantly low: t0: 0.80, t1: 0.76, t2: 0.108 és t3: 0.90. However, if the same factor is taken into account between the tests (see Table 3.), we may conclude that the output test results can be more accurately predicted in comparison with the pre-test. The correlation factor between the output tests shows a high value.

TABLE III. CORRELATIONS

		t0	t1	t2	t3
t0	Pearson Correlation	1	.422**	.451**	.488**
	N	1076	730	717	674
t1	Pearson Correlation	.422**	1	.479**	.519**
	N	730	730	689	643
t2	Pearson Correlation	.451**	.479**	1	.531**
	N	717	689	717	654
t3	Pearson Correlation	.488**	.519**	.531**	1
	N	674	643	654	674

We can draw a conclusion, that those participants who scored well at the pre-test stage had expectedly a better achievement during the final testing stage. As a proof, we can look at the correlation factor between the pre-test and the end tests which is 0.545. The positive values of the correlation factor reflect this idea that higher user activities (mouse movement, video play) can be related to better test results, though these correlations are considerably weaker by comparing them to the pre-tests.

If we compare the average outcome of the three output tests with the pre-tests, the R Square statistical value is 0.297 which shows a strong correlation.

We created 4 categories from the cities where the participants come from. The correlation coefficients reveal that there is no connection between the size of a town and the number of mouse clicks, video play, or mouse movement, however, as the size of a town grows the achievement results are better at all four tests (t0: 0.104. t1: 0.94, t2: 0.163 és t3: 0.188).

Learners' age and points scored at the tests show a correlation coefficient of 0.279 which demonstrates that age plays a significant role at the test results, namely older participants had better results and are more proficient regarding course topics, as revealed during pre-testing (correlation coefficient: 0.291).

If we create three clusters taking into account mouse movements, clicking, and test results, then the algorithm sorts those cases into the first cluster (n=279) who had few activities and had worse end term results (see Fig 3..). Remark: the values were normalized because of the clustering so the table shows negative results as well. The second cluster (n=102) contains those participants who were active but had an average end term results. Finally, the third cluster (n=251) demonstrates those learners who despite the fact that had less mouse movements or clickings achieved a good result.

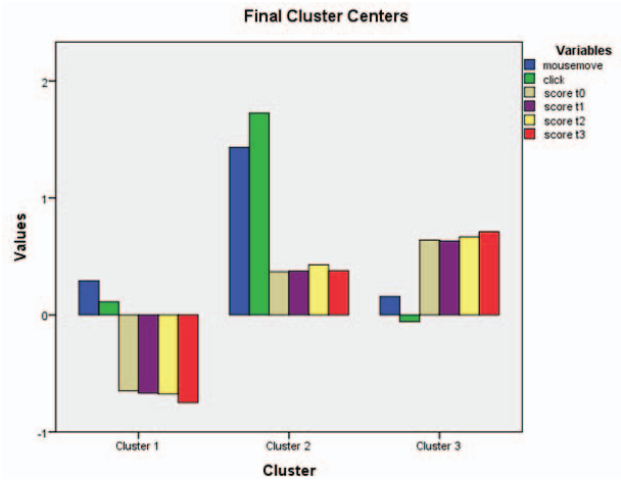

Fig. 4. Clustering depending on mouse movements, clicking and test results

To have a better visual overview, we show the clusters in a scatter diagram too (see Fig. 4.).

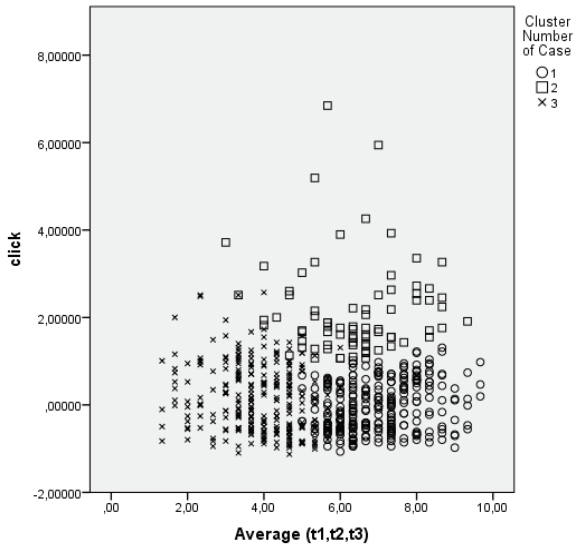

Fig. 5. Clustering shown on scatter diagram

V. CONCLUSION

By processing the data which was recorded by the self-developed logging module, it can be determined that those participants who fill in every test were more active during the course (more mouse movements, scrolling etc.), and the viewing of the videos positively influenced the completion of the course. Female participants had better scores on every test, and also more of them had successfully completed the course. According to the values of correlation coefficients and R Square Statistical value it can be stated, that those participants who scored well at the pre-test stage expectedly had better results during the final testing stage, so output test results can be more accurately predicted. The age of students and the size of their hometown had also influenced the outcome of the tests, namely: the older participants gained better results, and also better scores had been achieved by students living in bigger cities. By creating clusters, taking into account mouse movements, clicking, and test results, the first cluster belongs to students with less activities and worse scores. The second cluster contains those participants who were active but had an average end term result. The third cluster demonstrates students with less mouse movements or clicks, but with a good test outcome.

The collected data has been processed using statistical methods and basic data mining techniques in order to get a clearer picture of learning habits of the 5,6,7 and 8 grade primary school students from Vojvodina. Further plans are to apply the Statistical Matching method to predict student achievements.

REFERENCES

[1] Allen, I. E., & Seaman, J. (2008). Staying the course: Online education in the United States 2008. Needham, MA: The Sloan Consortium. URL: https://www.onlinelearningsurvey.com/reports/staying-the-course.pdf (2017.11.10.)

[2] T. Butler, M. Haldeman, E. Laurans: Creating Sound Policy for Digital Learning. [online] Washington: Thomas B. Fordham Institute, 2012.01.11.
URL:http://www.edexcellencemedia.net/publications/2012/20120110-the-costs-of-online-learning/20120110-the-costs-of-online-learning.pdf (2018.01.20)

[3] P. Esztelecki - G. Kőrösi (2015): Idegennyelv-tanulás megvalósítása online eszközökkel. 6. Báthory-Brassai nemzetközi konferencia.URL: http://www.bbk.alfanet.eu/userspace/6bbk2015_minden/6BBK2015_Tanulmany_kotetek/6BBK_konyv-2.pdf (2018.02.01)

[4] A. Hershkovitz , R. Nachmias (2011). Online persistence in higher education web-supported courses. Internet and Higher Education 14 (2011) 98–106.

[5] Boyd, D., & Crawford, K. (2012). Critical questions for Big Data. Information, Communication & Society, 15(5), 662—679. DOI: 10.1080/1369118X.2012.678878

[6] Rijmenam van M. (2013). Big Data Will Revolutionize Learning. Smart Data Collective.
URL:http://www.smartdatacollective.com/bigdatastartups/121261/big-data-will-revolutionize-learning (2018.02.01.)

[7] C. Romero and S. Ventura (2010). Educational Data Mining: A Review of the State of the Art," IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 40, no. 6, pp. 601–618, Nov. 2010..

[8] Carr, S. (2000). As distance education comes of age, the challenge is keeping the students. The Chronicle of Higher Education, 46(23), 39−41.

[9] J. Park (2007). Factors related to learner dropout in online learning. In: Nafukho, F. M., Chermack, T. H., & Graham, C. M. (szerk.): Proceedings of the Academy of Human Resource Development. Indianapolis: Annual Conference, 2007. 1–8 .p

Clickstream-based outcome prediction in short video MOOCs

Gábor Kőrösi
Institute of Informatics
University of Szeged
korosig@inf.u-szeged.hu

Péter Esztelecki
Institute of Informatics
University of Szeged
epeter@inf.u-szeged.hu

Richard Farkas
HAS Research Group on
Artificial Intelligence,
University of Szeged
rfarkas@inf.u-szeged.hu

Krisztina Tóth
TBA21 Ltd.
ktoth@tba21.hu

***Abstract*— In this paper, we present a data mining approach for analysing students' clickstream data logged by an e-learning platform and we propose a machine learning procedure to predict course completion of students. For this, we used data from a short MOOC course which was motivated by the teachers of elementary schools. We show that machine learning approaches can accurately predict the course outcome based on clickstream data and also highlight patterns in data which provide useful insights to MOOC developers.**

Keywords*— *data mining, clickstream, edm, predicting

I. Introduction

Over the past decade, technology and the internet have had a huge impact on the ways we learn. One of the most used parts of these innovations is the Massive Open Online Course (MOOC). Despite their early promise, however, MOOCs are still relatively unexplored and poorly understood [2]. Meanwhile, MOOCs often attract an enormous number of registrants, but only a small fraction of them can successfully complete their courses. Even of those students who declare at the start of a course an intent to complete it, 75% do not do so[6]. These high drop-off rates are often attributed to factors such as low teacher-to-student ratios, the asynchronous nature of interaction, and heterogeneous educational backgrounds and motivations, which make it difficult to scale the efficacy of traditional teaching methods with the size of the student body [4] [14]. Several researchers have analysed the server logs associated with these MOOCs to determine the factors associated with students dropping out, such as [5] [14] and conclude that student dropout rates are a major deterrent to the growth and success of MOOCs [6] [7]. [5][8]. As more and more higher education institutions make their courses available for learners through platforms such as MOOCs, the immense amount of data generated make it possible to provide continuous and automated assessment of student progress [9].

Analysing MOOC server log data in order to identify student drop-out patterns is an Educational Data Mining (EDM) task. Data is recorded during the time when learners are interacting with the MOOC platform providing a unique opportunity to learn about the efficacy of different resources, build predictive models that can help develop interventions and propose/recommend strategies for the learner [11]. By detecting whether student behaviour changes in a significant manner over the time-period of a particular term, we could identify students who increase, decrease, or show no changes in their clickstream activities, and whether these changes relate to course performance [3]. Student clickstream data has been the subject of a number of prior studies, such as the investigation of potential predictive relationships between online student activity and student outcomes, such as course grades.

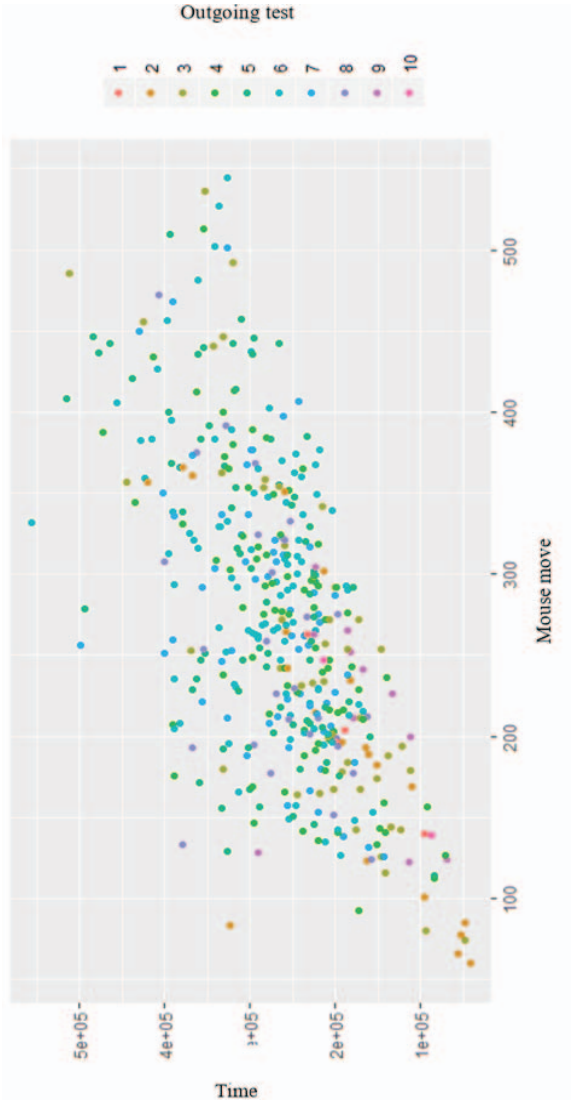

Fig. 1. Student achievement by behavior

There are two main MOOCs which have been investigated. The first group uses a huge log file from edX or Coursera [5,6,10,12,14] which has been generated by big Universities, such as Stanford or MIT. This data has been generated by ten or

978-1-5386-4599-4/18/$31.00 ©2018 IEEE

more thousand self-motivated students. On the other side, there is a second group which wants to create a successful prediction model from school-class-related e-learning platforms [1,2,37,,12,,,15] based on very shallow data from Moodle or Moodle-based forums. In our research, we analyse the log-data of students who were motivated by their teacher and their school to attend and complete the short (few- day-long) MOOC course. Our work has got similarities with both research avenues. Our logged data is very similar - wide and deep – to the data from edX and Coursera, even though it had been created in a much shorter period than that and is of a school-class nature. This log file gave us an opportunity to study clickstream data and user attitudes in short MOOC's. In this study, we present classification models that utilize data about the activities of students in courses to predict their final exam outcome. We propose a feature space of 263 attributes to describe students' clickstream data. Then, we apply various feature selection and various classification approaches.

The main contributions of our investigation are that our data mining procedure is able to accurately predict the success of students even if using short MOOC courses, and we highlight features which influence the classifier results the most, hence providing useful insights for MOOC developers.

II. DATASET

In this paper, we focus on clickstream data from a course which was recorded by an E-learning platform in the 2016-2017 academic year. In the course of recording, clickstream data was obtained through our course management system in the form of student IDs, time stamps, and activities.

The samples of data are constructed form a course named TÉBIA which involved upper grade pupils form 20 elementary schools. Components of a previously used and tested learning material were taken as the basis of the course content, which included an initial test with a video lesson and 3 further units. Every unit consisted of an obligatory video task and further optional textual learning material. To complete a unit, students had to solve 3 tests with a minimum score of 5 points out of 10. Every unit ended in a test with a maximum score of 10 points, except for the initial test. The structure of the learning material is demonstrated in Table I.

TABLE I. COURSE CONTENTS

Course name	TÉBIA
Content	Basics of Conscious and Safe Internet Usage
Time frame	6 weeks
Parts of the Learning Material	Introduction:Video (3.37 min., Embed);
	Digital footprint: Video (14.04 min, Embed);HTML embedded text;
	Conscious and Safe Internet Usage: Video (13.07 min, Embed); HTML embedded text; External link;
	Online bullying: Video(13.31 min, Embed); HTML embedded text;;Extra video (11.55 min, Embed);

Fig. 2. Distribution of students' final scores (n=603)

The distribution of students' final scores shows a Gaussian distribution (Fig. 2) which supports the validity of the outcome test.

A. Data cleaning

The types of activities recorded are those which correspond to broad categories of student behaviour, such as previewing lectures, mouse behaviours (move, scroll, click), video watching attitudes and text inputs. For instance, the course we examine in this paper had 1370 registered students who generated 2.430.975 click events over a 6-week period. The portal recorded 1370 students and lecturers, out of which only 1077 filled in and completed the initial test (Q0). As Fig 1 shows, the noisy and complex nature of this set of data made it impossible to use simple statistical or clustering methods to create a predictive model. Those students who had an output test but had insufficient amount of activities were eliminated from the measurement. The number of obtained results amounted to 603. According to conditions set to complete the course, we split the group (Q1>=5 and Q2>=5 and, Q3>=5) into two parts, which were labeled as 0 ("Failed") and 1 ("Completed").

B. Preliminary investigation

To investigate the structure of the data and understand user behaviour, we visualized the class-labelwise distributions of several log properties.. Because of the unbalanced nature of the data (n "Failed" = 419, n "Completed" = 184) we present density distribution. The following density figures show the differences between failed and completed students' main attributes. The final diagrams show a significant overlap between the two groups, which makes it harder to adjust the weighting settings. Without striving to present an extensive number of differences between the two groups, we show some of them in the following tables and figures. (Fig. 3, Fig. 4, Fig. 5, Table II, Table III, Table IV) All the following diagrams were constructed using the Ggpolt R package of Wickham, et al.[16]

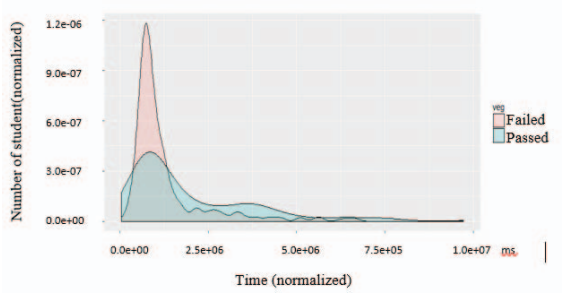

Fig. 3. Time spent in course

TABLE II. TIME SPENT IN COURSE

„Failed”

Min.	1st Qu.	Median	Mean	3rdQu.	Max.	NA's
51000	684300	861900	1294000	1291000	9725000	3

„ Completed

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
123500	750400	1098000	1987000	2970000	9462000	3

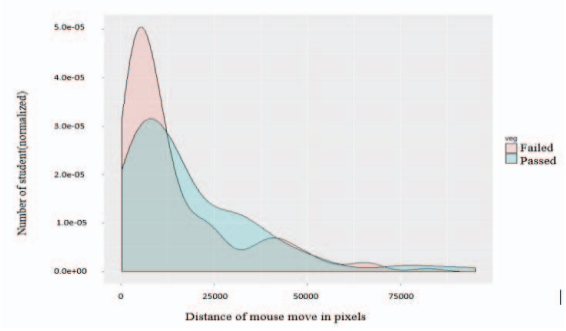

Fig. 4. Average distance of mouse in course contents

TABLE III. AVERAGE DISTANCE OF MOUSE IN COURSE CONTENTS

„Failed”

Min.	1st Qu.	Median	Mean	3rdQu.	Max.	NA's
226	4325	8190	14800	20570	82210	262

„ Completed”

Min.	1st Qu.	Median	Mean	3rdQu.	Max.	NA's
451	5632	12990	19040	28520	95030	76

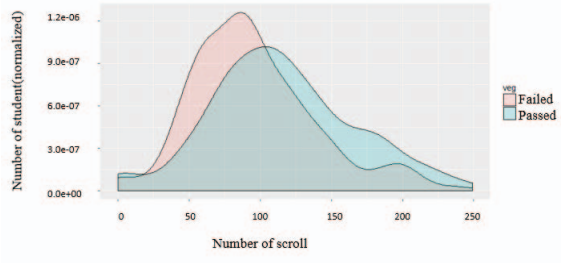

Fig. 5. Average number of scrolls in course contents

TABLE IV. AVERAGE NUMBER OF SRCOLLS IN COURSE CONTENTS

„Failed”

Min.	1st Qu.	Median	Mean	3rdQu.	Max.	NA's
0	64	89.50	95.58	118.80	244.00	9

„ Completed”

Min.	1st Qu.	Median	Mean	3rdQu.	Max.	NA's
0	81	109	115	144	249	9

III. MACHINE LEARNING EXPERIMENTS

We carried out machine learning experiments using clickstream log-data to predict whether a particular student will fail or succeed in the final exam of the MOOC. We employed the Rminer [17] package of R.

A. Feature space

We defined 263 features to describe our clickstream data. There were two types of data. In the first group, there was the data which was collected during the filling process in the incoming test. The second type was the clickstream which was collected during the learning process in the three parts of the curriculum (see Table 1). This collection was divided into 18 main categories: binary and numeral answers provided to input and output tests (28+60), time spent on the quizzes (6) and the sites of the curriculum (7), the number of visits to the site of quizzes (6) and site of curriculum (7), the mouse move distance in pixels (13), the average mouse speed (13), the cumulated data (6), the number of mouse movement on a page (13), the number of clicks on a page (13), the use of test buttons during the input/output testing (4), the number of scrolls on a page (13), the last login date to a page compared to the first login to the site (13), the first login date compared to the first login to the site (13), the days spent on the sites (13), the mean behaviour on the sites (19), the number of calendar days between the output tests (7), binary output results (1), output results (2), user-related data (6).B. Feature selection

We investigated different feature selection methods, and the gain-ration function in the FSelector package [15] proved to be the most effective.

Gain-ratio examines all the parameters one-by-one and creates a hierarchy, which distinguishes weak and strong correlational connections. The FSelector package was designed to handle such problems and the most useful functions of all were the chi.square and gain.ratio filtering algorithms. Between the two, the latter provided accurate calculations so the choice to present the underlying theory.

The information gain method chooses a split based on which attribute provides the greatest information gain. The gain is measured in bits. Although this method provides satisfactory results, it favours splitting on variables that have many attributes. The information gain ratio method incorporates the value of a split to determine what proportion of the information gain is valuable for that split. The split with the greatest information gain ratio is chosen. [13] The information gain calculation starts by determining the information of the training data. The information in a response value, r, is calculated in the following expression:

$$-log_2\left(\frac{freq(r,T)}{|T|}\right)$$

T represents the training data and $|T|$ is the number of observations. To determine the expected information of the training data, sum this expression for every possible response value:

$$I(T) = -\sum_{i=1}^{n} \frac{freq(r_i,T)}{|T|} \times log_2\left(\frac{freq(r_i,T}{|T|}\right)$$

Here, n is the total number of response values. This value is also referred to as the *entropy* of the training data.

Next, consider a split S on a variable X with m possible attributes. The expected information provided by that split is calculated by the following equation:

$$I_s(T) = \sum_{j=1}^{m} \frac{|T_j|}{|T|} \times I(T_j)$$

In this equation, T_j represents the observations that contain the j^{th} attribute.

The information gain of split S is calculated by the following equation:

$$G(S) = I(S) - I_s(T)$$

Information gain ratio attempts to correct the information gain calculation by introducing a split information value. The split information is calculated by the following equation:

$$SI(S) = -\sum_{j=1}^{m} \frac{|T_j|}{|T|} \times log_2\left(\frac{|T_j|}{|T|}\right)$$

As its name suggests, the information gain ratio is the ratio of the information gain to the split information:

$$GR(S) = \frac{G(S)}{SI(S)}$$

B. *Prediction Models*

Classifying whether the student failed or completed the course was the core goal of this study. We train various machine learning models for prediction. Because of the limited size of our dataset, we applied the LEAVE-ONE-OUT cross validation method.

We comparatively experimented with the following classifiers: "lr"- logistic regression, "xgboost" - eXtreme Gradient Boosting ,"mlpe"- multilayer perceptron ensemble, "mlp"- multilayer perceptron with one hidden layer ,"ksvm" - support vector machine, ,"kknn"- k-nearest neighbor,"naiveBayes"- naive bayes,"naive","ctree"- conditional inference tree, "rpart"- – decision tree,"randomForest"- random forest algorithm,"boosting"- boosting ,"bagging"- bagging.

IV. Experimental Results

During the data cleaning process, we reduced the number of students from 1370 to 603. The preliminary results showed that every student-user had a unique click stream pattern, which was very similar and independent of user achievement and final scores. Such a finding underpinned that data saved by the MOOC system is suitable to build prediction models. It will be possible to help educational institutions to fight to lower the drop-out rate. They could also take action to help users whose achievement results fall below the average to prevent negative outcomes.

We carried out binary classification experiments to predict whether a student will successfully complete the MOOC and get the certificate. There were 429 students out of 603 who successfully completed the course, i.e. the most frequent class baseline is 71%.

The gain-ratio feature selection ranked features in an ascending order and the experiment showed that approximately the top 60 features are useful. Results were completely tested in 60 cases and by halving further 30, 15. The following table summarizes accuracies achieved by the 12 classifiers using the top 60 features. Besides accuracy (ACC), we also report the recall, precision, and F-score values of the Completed class.

Figure 6 provides an overview of the classifiers using only the top 15, 30 and 60 features. We can conclude that along the 30 properties, the most accurate results were achieved by the supported vector machine and the random forest function. While in the case of 60 features, the most accurate was the bagging function. In the end, we could say that the most successful model were the bagging (ACC 80.10%) and the random forest (ACC 79.44%) methods (Table V., Fig. 6.).

TABLE V. Performance of Certificate Earner Prediction with Different Methods (%), the most weighted 60 feature

	Bagging	Boosting	Ctree	Kknn	Ksvm	Lr	Mlpe	Random Forest
ACC	80.1	78.11	64.34	71.14	77.61	78.44	73.47	79.44
RECALL	91.14	87.88	79.25	76.92	94.87	87.65	82.05	92.07
PRECISION	82.66	82.49	72.96	81.48	78.27	83	80.92	81.44
F1	86.7	85.1	75.98	79.14	85.77	85.26	81.48	86.43

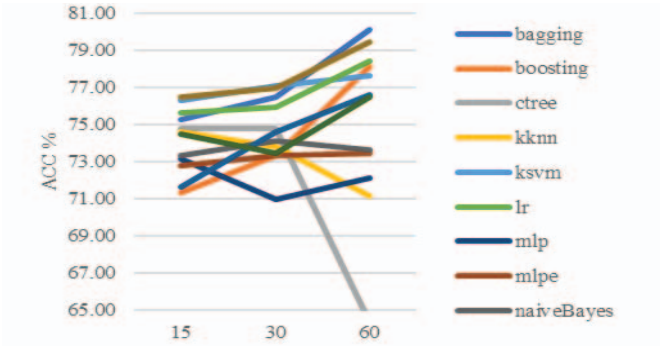

Fig. 6. Average Prediction performance in the function of the number of features for training

V. Discussion

This paper describes a statistical methodology for predicting binary outcome in a set of data which was created in short MOOC and driven by the teacher. Based on these data sets, we found a successful model which was influenced by a couple of strong features. The accuracy of the models has achieved satisfactory accuracy of more than 80%. It confirms the supposition that we are able to efficiently predict learning outcomes. Through more detailed research, the two models show significant differences. The best results were achieved by those features which were connected to the learning material or the average value of cursor distance on a curriculum page. Based on our methods, we could describe which were the most notable features in our prediction models. As we can see in Fig 7, the most highly weighted features in feature selection process over data stream. As we expected, the highest weight got the input test grades, after which followed the average time, mouse speed and mouse distance spent in the whole course. The other important things were the number of clicks, and scrolls, and the number of mouse moves on the page of the curriculum. At the beginning, we expected the amount of time would most influence the outcome because those who spend more time on the system, would learn more. In the end, we realized that taking more time in the course does not have a considerable effect on the outcome of grades. On the other hand, as in ordinary

schools, the number of days spent learning and testing has shown its effect during the evaluation process.

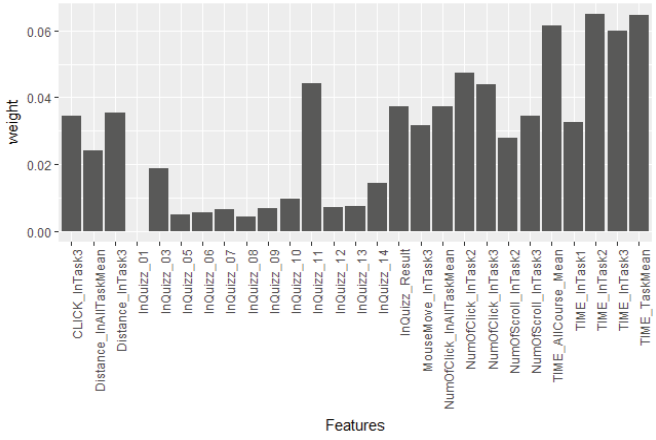

Fig. 7. The highest weighted features

VI. CONCLUSION

Student clickstream data is inherently difficult to work with given its complex and noisy nature [3]. Several data mining applications are focused on educators, where the object is to help create accurate feedback, categorization of learners based on their abilities, course creation, and instructional plans.

This paper introduced a machine learning methodology for outcome classification of short video MOOCs based on clickstream data. Our primary goal was to do binary prediction of course completion and of student engagement. Our models could predict who would "Fail" or "Complete" an online course, which would be an immense help for the faculties that provide e-learning courses. Despite a relatively low sample size, we could still render click stream based predictive algorithms. We proposed 263 features to describe clickstream data of short video MOOCs. We employed feature selection and binary classification techniques in a leave-one-out cross validation evaluation setting. The most efficient tools for our models were the Random Forest and Bagging achieving with approximately 80% accuracy.

While the results in this paper are promising and there are interesting methodological avenues to pursue, the most important future direction from an education research perspective will involve more in-depth investigation of the utility of these types of methods in terms of providing actionable insights that are relevant to the practice of education.

ACKNOWLEDGEMENT

The project was supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002).

REFERENCES

[1] C. Robinson, M. Yeomans, J. Reich, C. Hulleman, and H. Gehlbach, "Forecasting Student Achievement in MOOCs with Natural Language Processing", *Sixth International Conference on Learning Analytics & Knowledge*, Universitz of Edinburhg, Edinburgh, pp. 383–387, 2016.

[2] Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Engaging with Massive Online Courses", *23rd international conference on World wide web*, Seoul, Korea, pp. 687–698, April 7–11, 2014.

[3] J. Park, K. Denaro, F. Rodriguez, P. Smyth, M. Warschauer, "Detecting Changes in Student Behavior from Clickstream Data", *Seventh International Conference on Learning Analytics & Knowledge*, Vancouver, BC, Canada pp. 21–30, 2017.

[4] X. Wang, D. Yang, M. Wen, K. R. Koedinger, and C. P. Rose, "Investigating how student's cognitive behavior in MOOC discussion forum a ect learning gains". *In Proceedings of the EDM Conference*,International Educational Data Mining Society (IEDMS), pp. 226–233, 2015.

[5] Z. Ren, H. Rangwala, and A. Johri, "Predicting Performance on MOOC Assessments using Multi-Regression Models", *Computers and Society*, 2016.

[6] H. Daumé, D. Goldwasser, L. Getoor, B. Huang, and A. Ramesh, "Modeling Learner Engagement in MOOCs using Probabilistic Soft Logic", 2013.

[7] U. Anderson, T. Arvemo, and M. Gellerstedt, "Can Measurements of Online Behavior Predict Course Performance?", *7th International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC 2016 and the 7th International Multi-Conference onSociety and Information Technologies: ICSIT 2016*: Volume II (Post-Conference Edition), pp. 4–9, 2016.

[8] S. Tang, J. C. Peterson, and Z. A. Pardos, "Modelling Student Behavior using Granular Large Scale Action Data from a MOOC", 2016.

[9] M. M. Ashenafi, G. Riccardi, M. Ronchetti, "Predicting Students' Final Exam Scores from their Course Activities*", Frontiers in Education Conference (FIE),* 2015 pp. 10–22, 2015.

[10] C. G. Brinton, S. Buccapatnam, M. Chiang, and H. V. Poor, "Mining MOOC Clickstreams: On the Relationship Between Learner Behavior and Performance", Cornel University, 2015.

[11] S. Boyer, and K. Veeramachaneni, "Transfer Learning for Predictive Models in Massive Open Online Courses" , In: Conati C., Heffernan N., Mitrovic A., Verdejo M. (eds) *Artificial Intelligence in Education. AIED 2015.* Lecture Notes in Computer Science, vol. 9112. Springer, Cham, 2015.

[12] Ch. G. Brinton, and M. Chiang, "MOOC performance prediction via clickstream data and social learning networks", *Computer Communications (INFOCOM)*, 2015 IEEE Conference , 2015.

[13] Inc. 2015. SAS® Visual Analytics 7.2: User's Guide. Cary, NC: SAS Institute Inc.

[14] Y. Tsung-Yen, Ch. G. Brinton, C. Joe-Wong, "Behavior-Based Grade Prediction for MOOCs via Time Series Neural Networks", IEEE Journal of Selected Topics in Signal Processing ,vol. 11, Issue: 5, Aug. 2017

[15] P. Romanski, L. Kotthoff, Package 'FSelector', 2016 , related:https://CRAN.R-project.org/package=Fselector

[16] H. Wickham, W. Chang, "Create Elegant Data Visualisations Using the Grammar of Graphics" Package 'GGplot', 2016 , related: https://CRAN.R-project.org/package=ggplot2

[17] P. Cortez, "Data Mining Classification and Regression Methods", Rminer package, 2016 , related: https://CRAN.R-project.org/package=rminer