Detlef Kromker und Ulrik Schroeder (Hrsg.): Die 16. E-Learning Fachtagung Informatik, Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2018

Student Success Prediction and the Trade-Off between Big Data and Data Minimization

Hendrik Heuer[1] Andreas Breiter[2]

Abstract: This paper explores student's daily activity in a virtual learning environment in the anonymized Open University Learning Analytics Dataset (OULAD). We show that the daily activity of students can be used to predict their success, i.e. whether they pass or fail a course, with high accuracy. This is important since daily activity can be easily obtained and anonymized. To support this, we show that the binary information whether a student was active on a given day has similar predictive power as a combination of the exact number of clicks on the given day and sensitive private data like gender, disability, and highest educational level. We further show that the anonymized activity data can be used to group students. We identify different student types based on their daily binarized activity and outline how educators and system developers can utilize this to address different learning types. Our primary stakeholders are designers and developers of learning analytics systems as well as those who commission such systems. We discuss the privacy and design implications of our findings for data mining in educational contexts against the background of the principle of data minimization and the General Data Protection Regulation (GDPR) of the European Union.

Keywords: Learning analytics, MOOCs, daily activity, machine learning, data science, group formation, digital traces, privacy, clickstream, student data, student performance.

1 Introduction

Virtual learning environments (VLEs) are tools to support distance education as well as classroom teaching. They aim to enhance the learner-teacher interaction. Moreover, their logs offer unprecedented access to student activity. In this paper, we use student's daily activity in a virtual learning environment to predict their final result for a class. We also show daily activity can be used to form groups of students. Our work is motivated by the Janus-faced character of such systems: on the one hand, they do offer the potential to help students study and teachers teach, on the other hand, they are tantamount to total surveillance with its inherent privacy risks. We aim to mitigate this by finding a good trade-off between big data and the principle of data minimization. We explore what data is minimally necessary to predict whether a student of a VLE is at risk of failing a class. This motivation connects to the General Data Protection Regulation (GDPR) of the European Union [Eu16]. In Article 5 1. (c), the GDPR describes the Principle of Data Minimization, which states that personal data shall be "limited to what is necessary in relation to the purposes for which they are processed". The GDPR was adopted in April 2016 and became enforceable in May 2018.

[1] University of Bremen, Institute for Information Management Bremen (ifib) and Centre for Media, Communication and Information Research (ZeMKI), Am Fallturm 1, 28359 Bremen, hheuer@ifib.de

[2] University of Bremen, Institute for Information Management Bremen (ifib) and Centre for Media, Communication and Information Research (ZeMKI), Am Fallturm 1, 28359 Bremen, abreiter@ifib.de

GDPR protects the privacy rights of people in Europe regardless of where their data is collected. This extraterritorial nature of the GDPR makes it the de facto standard for privacy around the world. Motivated by the GDPR and the Principle of Data Minimization, our paper contributes to the applicability of data science in an educational context by answering the following questions: 1. Can data about the daily activity of students be used to predict the success of a student in a virtual learning environment? 2. How fine-grained does the data about daily activity need to be? and 3. What kind of student types can be derived from student's daily activity?

2 Related Work

Our research is positioned in the emerging field of learning analytics (LA) and educational data mining (EDM). Papamitsiou and Economides provide an overview of empirical evidence behind key objectives of the potential adoption of LA/EDM in generic educational strategic planning [PE14]. For this, they examined experimental case studies between 2008 and 2013 and identified 209 mature pieces of research work in this area. They identified five distinct directions of the LA/EDM empirical research: 1. modeling of students and their behavior, 2. prediction of performance, 3. increasing (self-)reflection and (self-)awareness, 4. prediction of dropout and retention, and 5. recommendation of resources. In this paper, we focus on the prediction of performance and the prediction of dropout and retention. Our primary stakeholders are designers and developers of learning analytics systems as well as those who commission such systems. Slade et al. state that learning analytics has the potential to enable higher education institutions to increase their understanding of their students' learning needs, but point out serious ethical challenges of big data [SP13]. These include the issues of the location and interpretation of data, informed consent, privacy and the de-identification of data, and the classification and management of data. Our work connects to both the interpretation of the data and the privacy and de-identification aspects. Lukarov et al. divided the goals of learning analytics into three groups: 1) those that inform the design of learning analytics tools, 2) those that involve a behavioral reaction of the teacher, and 3) those that involve a behavioral reaction of the student [LCS]. In this paper, we analyze the behavior of students to inform the design of learning analytics tools. For this, we focus on the daily activities in a virtual learning environment and machine learning techniques. Chatti et al. provide a reference model for learning analytics that distinguishes four dimensions: data and environments (what?), stakeholders (who?), objectives (why?), and methods (how?) [Ch12]. In this paper, we show that with a clear why, i.e. predicting student success, the what and how can be adapted to fit in a data minimization framework in line with GDPR.

To predict student success, we apply machine learning. Mitchell defined machine learning as a computer program that learns “from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." [MCM13]. In this paper, we use machine learning to accomplish two tasks: 1. a supervised classification task, where we predict whether students pass or fail a course and 2. an unsupervised clustering task, where we group students based on their daily activities. Machine learning is an area where more data is commonly equated with better predictions.

In this paper, we show that feature engineering can enable practitioners to predict student's success in a way that is consistent with the principle of data minimization. This connects to the fundamental ethical issues and dilemmas of learning analytics [SP13]. Like with other learning analytics tasks, we have to pay special attention to class imbalances and use appropriate evaluation metrics to avoid being misled by the scores of the system [HZZ17].

Massive open online courses (MOOCs) have become increasingly popular in the last years, with many courses reporting large numbers of enrolled students. At the same time, retention and pass rates have been very low with many MOOCs [Ke15, He15]. This motivated a large body of research that tries to predict student success, performance, and retention. A large number of participants and the fact that the majority, if not all interactions between teacher and students are mediated by the platform, generates large amounts of data well-suited for machine learning algorithms. The main motivation for this line of research is to improve retention of students [Wo13] and offer them help at the right time, i.e. when they are about to drop out [Wo14b]. While this became increasingly important with the advent of Massive Open Online Courses (MOOCs) [He15], it is relevant to other distance learning institutions as well [Wo14b]. In addition to that, higher education institutions and K-12 schools all over the world are exploring techniques like “flipped or inverted classroom models" or blended learning [Fu12, HHG15]. For all such approaches, it is important to understand what factors contribute to the success of students in a particular learning setting. A large body of prior research aimed at predicting students that are at-risk of failing their classes exists [Wo13, HZZ17, Wo14b, Ku15, He15]. In many cases, predictive models are constructed from legacy data using machine learning methods to predict student behavior and their learning paths. For the prediction, demographic information, clickstream data, as well as formal assessments can be used [BH18]. This connects to the digital traces, i.e. numerically produced correlations of disparate kinds of data, inevitably generated by our use of media [HBF18]. Breiter and Hepp investigated the challenges of putting digital traces in context. For this, they use data from learning management systems as an example and outline strategies of how to put this data into context using qualitative methods [BH18]. For this paper, we focus on quantitative data. In a dataset from the Open University in the United Kingdom, one of the largest distance learning institutions in the world [Wo13]. Student success can be predicted e.g. by comparing it to their own previous behavior, based on students with similar learning behavior or using data from the current course [Wo13, HZZ17]. This paper connects to prior work on predicting at-risk students from clicking behavior in a virtual learning environment [Wo13]. Kennedy et al. investigated whether learners' prior knowledge, skills, and activities influence MOOC performance [Ke15]. They found that prior knowledge is the most significant predictor of MOOC success followed by students' ability to revise. Kizilcec et al. showed that individuals with strong self-regulated learning (SRL) skills, can learn faster and outperform those with weaker SRL skills [KPSM17]. They also found that several learner characteristics, including demographics and motivation, predicted learners' SRL skills.

We also connect to prior research that investigated the impact of Learning Analytics (LA) and Educational Data Mining (EDM) on adaptive learning [PE14]. Kizilcec et al. grouped learners based on their patterns of interaction with video lectures and assessments and used this to develop adaptive course features for particular subpopulations of learners [KPS13].

In this paper, we devise a scheme to cluster students based on their learning activities and address how this can be used by teachers and system developers alike. For this, we cluster users based on their daily activities, which can be applied to form groups and connects to prior work [Ko13, Be17]. Bellhäuser et al. examined the effect of personality traits on satisfaction and student performance and showed that a mix of extroverted and introverted students is beneficial for the performance of groups [Be17]. Mixing students based on their diligence (German: *Gewissenhaftigkeit*) was also found to be beneficial. Here again, educational data mining offers powerful new ways to support teacher's teaching and student's learning, which needs to be reconciled with a jurisdiction aimed to protect the privacy of its citizens.

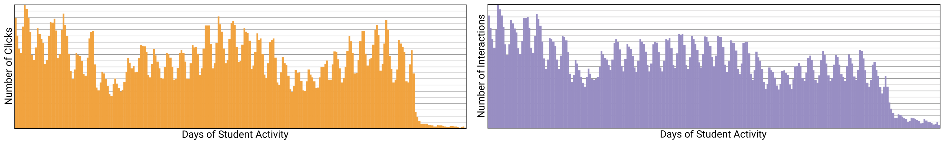
3 Methodology

We used the anonymized Open University Learning Analytics Dataset (OULAD) for our investigation [KHZ17]. The dataset was collected and analyzed at the Open University to provide informed guidance and to optimize learning materials. It consists of student information (demographic data), assessment results (performance data) and click-stream data of 22 courses and 32,593 students (learning activities data, i.e. the digital representation of their movements within a system). The click-stream data contains logs of all interactions within the virtual learning environment (VLE) of Open University, aggregated as daily summaries of student clicks (10,655,280 entries). Each data point has a student id, i.e. a unique identification number for the student, an identification code of the module, and an identification code of time the student registered for a module. We removed all data points with missing data. Out of the 32,593 students and their results, this yields 21,562 student-course combinations, out of which 6906 failed (32.03%) and 14656 passed (67.97%). The virtual learning environment contains information about the available materials and the student's interactions. Data includes the type of activity as well as the period of time when the assessment was supposed to be used. For each student and for each day of the semester, the number of clicks of the students is provided. This entails all interactions with the course material on a given day. For our exploration, we combined all activity types into a single metric, i.e. daily activity. This means that for each student-course combination, we have two 245-dimensional vectors of the 245 days of activity: one that represents whether a student was active on a given day (the interactions vector) and one that has the precise number of clicks (the clicks vector). The dataset also includes a student's final result in a course and data on all assessments, i.e. the assessment type, the date, as well as the weight of the assessment. For each assessment, the student's scores are also provided. In this paper, we only focus on the final results and disregard the data on other assessments. The final results are indicated as "Withdrawn", "Fail", "Pass", and "Distinction". Considering the myriad reasons for withdrawing from a class, we excluded this data from our analysis. We further simplified the analysis by merging "Pass" and "Distinction" and ignore all data points that were banked, i.e. transferred from a previous semester. The demographic data contains a student's gender, age band (i.e. not the precise age), his or her highest education level on the entry to the module, and whether the student has declared a disability. The dataset further includes information on how many times the student has attempted a module and his or her total number of credits. Finally, the geographic region, where the student lived while taking the course and its Index of Multiple Deprivation is included, which combines 37 separate indicators like income, employment, and crime to measure the relative deprivation of small areas.

3.1 Machine Learning Model

This paper aims to investigate the trade-off between big data and data minimization. For the classification tasks, we relied on a subset of the approaches used by Wolff et al. [Wo14a]. We report the following four supervised machine learning systems to derive our classification models: 1. *Decision Tree*, which consists of simple decision rules inferred from the data features, 2. *Random Forest*, which fits a number of decision trees on different subsets of the dataset and uses averaging to improve the predictive accuracy, 3. *Logistic Regression*, a maximum entropy classifier commonly used as a benchmark, and 4. *Support Vector Machine*, an advanced machine learning algorithm that transforms the input into a suitable high-dimensional space using the kernel trick. The goal was not to provide a comprehensive evaluation of complex machine learning systems for the task, but rather to focus on the influence of data preparation and feature engineering. Therefore, we focused on the effect of data minimisation on common predictors and did not explore more complex models like Bayesian networks or feed-forward or recurrent neural networks, since they are more complex and harder to train, which would have affected the replicability of our results. We implemented all models using the Python machine learning library scikit-learn [Pe11]. We did not perform any hyperparameter optimization and relied on the default values of sklearn 0.19.1, which can be found in the documentation for reference and reproduction. For each machine learning model, we performed a 5-fold cross-validation. This means that we subdivided our dataset into five non-overlapping subsets and trained and evaluated independent models. We report the average of the five runs.

For the clustering tasks, we used the k-means clustering implement of scikit-learn [Pe11]. K-means partitions datasets into clusters of data points that are similar to each other. The clusters are optimized using expectation-maximization, minimizing the within-cluster sum of squared criterion. To evaluate our machine learning classification models, a variety of metrics exist. Such metrics quantify the number of true positive (t_p), true negative (t_n), false positive (f_p), and false negative (f_n) that the classifier produced. In our evaluation, we report accuracy ($\frac{t_p+t_n}{t_p+t_n+f_p+f_n}$), precision ($\frac{t_p}{t_p+f_p}$), recall ($\frac{t_p}{t_p+f_n}$), and F1 ($\frac{2*(Precision*Recall)}{(Precision+Recall)}$). As noted, imbalanced data exacerbates the training and evaluation of classification models since metrics like accuracy and recall become misleading [HZZ17]. If 95% of students pass a course, a machine learning system that would always and only predict students passing would achieve an accuracy of 95% and a recall of 95%. Therefore, the precision and the F1 metric as the harmonic mean of precision and recall are important to assess the quality of the prediction.



(a) Clicks per Day

(b) Interactions per Day

Figure 1: Daily activity of 21,562 students in the Open University Learning Analytics Dataset as (a) click counts and (b) interaction counts (the binary data whether a student was active on a day or not).

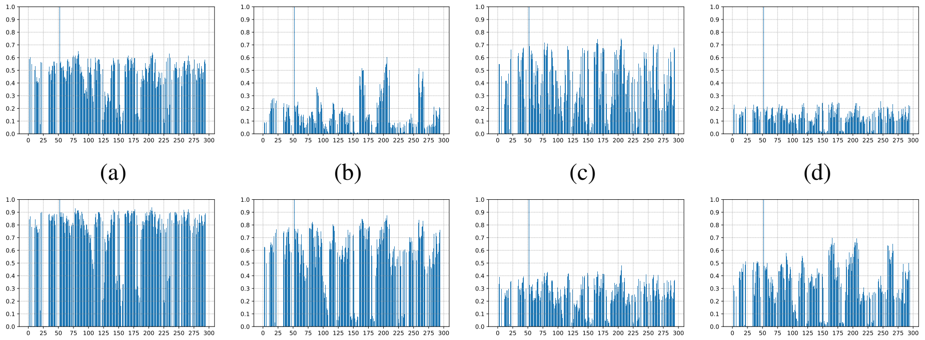
4 Results

During our initial exploratory data analysis, we noticed strong similarities of activity as measured by the clicks per day and the interactions per day that motivated our approach. Figure 1a visualizes the number of clicks per day as the sum of all students, Figure 1b aggregates all students that interacted with the virtual learning environment at all, i.e. those students that performed one or more clicks. The comparison of the two graphs shows that the patterns of interaction are similar. That said, the clicks per day in Figure 1a have some noteworthy peaks at the beginning, in the middle, and at the end of the course. Based on that, we trained a machine learning model to predict the success of a student in a university course from his or her daily activity in a virtual learning environment. For this, we applied four machine learning models (Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine) and evaluated them in a 5-fold validation based on accuracy, precision, recall, and F1. We compared three different feature sets: 1. *Student Info*: the predictive power of demographic student information like gender, region, and highest education, 2. *Daily Activity*: the daily activity in the virtual learning environment, and 3. *Student Info & Daily Activity*: a combination of demographic student information and daily activity in the virtual learning environment. For 2. and 3., we also differentiate how we represented daily activity: a. binary ($active = 1$, $else = 0$), b. the normalized count ($\frac{x_i}{max(x)}$), and c. the actual count. We removed all data points with missing data. Out of the 32,593 students and their results, this yields 21,562 student-course combinations, out of which 6906 failed (32.03%) and 14656 passed (67.97%). We repeat these numbers since they are important to contextualize and interpret our reported evaluation metrics. A model that would always predict that the student passes would achieve an accuracy of 67.97% and a recall of 67.97%. This means that this is the threshold for a model that makes any prediction at all (unlike in a balanced binary classification tasks, where this would be 50%).

Table 1 provides an overview of the results of these machine learning models. The task was to predict whether a student passes or fails a certain course. We see that *Student Info* alone does not allow any meaningful predictions. This is noteworthy since demographic data like a student's geographic location and socio-economic background are known to be a strong success indicator in K-12 [D'01, OE, Mu17]. *Daily Activity* in the virtual learning environment, on the other hand, has strong predictive power. Interestingly, we found that the binarized version, where we only indicate whether a student interacted with the learning environment at all, has stronger predictive power than the representation of the actual counts

Input	**Metric**	**Decision Tree**	**Random Forest**	**Logistic Regression**	**SVM (RBF)**
Student Info	F1	0.6703	0.7254	**0.7820**	0.7139
	P	0.6463	0.7516	0.8812	0.8367
	R	0.7019	0.7145	0.7171	0.7032
	ACC	0.5733	0.6255	**0.6762**	0.6310
Daily Activity (binary)	F1	0.8307	0.8997	0.8788	**0.9085**
	P	0.8005	0.9260	0.8662	0.9311
	R	0.8650	0.8770	0.9006	0.8896
	ACC	0.7791	0.8608	0.8427	**0.8739**
Daily Activity (normalized)	F1	0.8409	0.9031	0.8093	0.8093
	P	0.8675	0.8743	0.6797	0.6797
	R	0.8175	0.9357	1.0000	1.0000
	ACC	0.7906	0.8641	0.6797	0.6797
Daily Activity (counts)	F1	0.8392	0.9026	0.8485	0.8655
	P	0.8662	0.8740	0.8814	0.7651
	R	0.8156	0.9349	0.8405	0.9965
	ACC	0.7885	0.8635	0.8107	0.7894
Student Info & Daily Activity (binary)	F1	0.8251	0.9061	0.8867	**0.9132**
	P	0.7872	0.9372	0.8710	0.9333
	R	0.8694	0.8781	0.9052	0.8950
	ACC	0.7751	0.8685	0.8495	**0.8798**
Student Info & Daily Activity (normalized)	F1	0.8362	0.9070	0.8058	0.8093
	P	0.8091	0.9380	0.8925	0.9979
	R	0.8659	0.8792	0.7476	0.6806
	ACC	0.7849	0.8697	0.7144	0.6803
Student Info & Daily Activity (counts)	F1	0.8388	0.9052	0.8744	0.8669
	P	0.8664	0.8763	0.8879	0.7676
	R	0.8133	0.9373	0.8698	0.9958
	ACC	0.7877	0.8669	0.8351	0.7919

Table 1: Results for the machine learning task of predicting student success using demographic Student Info as well as different representations of Daily Activity as input.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2: Different student types based on their daily interactions (y-axis) per day (x-axis) identified via k-means clustering.

as a normalized number or as a total count. We found that a Support Vector Machine with a radial basis function as a kernel outperforms all other classifiers. The F1 score achieved in the 5-fold cross validation is 0.9085, the accuracy is 0.8739. Note that the *Random Forest* classifier is relatively close to this with F1 at 0.8997 and accuracy at 0.8608. Combining *Student Info & Daily Activity* yields a small but noticeable improvement. Here again, the *Support Vector Machine* outperforms all other classifiers with an F1 at 0.9132 and an accuracy at 0.8798, again with a *Random Forest* as runner-up with an F1 at 0.9061 and accuracy at 0.8685.

To find students who are similar to each other, we applied k-means clustering on the binary vector of daily activity, which gave us different student types. A cluster is a group of students that have a similar pattern of daily activity in the virtual learning environment. For each student-course combination, we used the binary 245-dimensional interactions vector. Based on an exploratory data analysis phase, we set k = 9. Figure 2 shows the different student types we discovered. We operationalize a student type as a pattern of daily activity shared across a large group of students. For each of the 9 clusters, we summed all interactions and divided by the number of students in the cluster. This yields a normalized overview of dates on which the students in the cluster interacted with the material. While we only present the aggregated graphs here, we also inspected the members of the cluster for the following analysis. Figure 2e is the cluster with the strongest activity. We can characterize the members of this cluster as power users. An analysis of the individual usage patterns shows long streaks, i.e. students that interact with the system every day. Figure 2i was omitted since it neither shows a visible pattern of specific dates nor are the individual members of the clusters of the cluster particularly active. Figure 2a, Figure 2c, Figure 2f, and Figure 2h are similar in their intensity and following a similar general pattern, though they are different in the distribution in streaks, e.g. comparing Figure 2c and Figure 2f.

5 Discussion

In this paper, we explore student's daily activity in a virtual learning environment with a special focus on the principle of data minimization. We found that daily activity is an important indicator of student performance and that daily activity can be used to predict whether a student will fail or not. We showed that in this particular context, daily activity is more important than highly personal information about the students. An in-depth review of the decision tree trained for the *Student Info & Daily Activity (binary)* input showed that for the five highest levels of the decision tree, the code module and the highest education of the student are the only demographic attributes that are not associated with the daily activity of the student. For this particular decision tree, the 201st day is the most predictive, followed by both the 138th and the 131st day. This means that in this context, a student's highest achieved education is the only non-activity based property necessary to predict whether the student would pass or fail. Other attributes like gender, age, disability, or the Index of Multiple Deprivation (IMB) of the place where the student lived, are less useful for the classification tasks than the information whether the student was active on a specific day or not.

Our findings shape ways of the educational transformation in two ways: 1. we show that such machine learning systems can help identify students at-risk and provide them with additional material to help them succeed and 2. this can be accomplished while adhering to the principle of data minimization. By only encoding whether a student was active or not, we simplified our data and model. We showed that private traits like gender, disability, and highest educational level are not needed for the success prediction task. As a result, we reduced the risk of discrimination based on such data. While the daily activity of clicks can be directly traced back to an individual, the binary representation only allows predicting them as groups (though it's of course still possible to narrow down the number of students that come into questions). This means that the binary representation yields a more powerful and more privacy-protective machine learning model, which is an important finding for future applications of machine learning in an e-learning context. Researchers and practitioners alike should prefer the simplest approach that minimizes necessary data. The k-means clustering yielded different patterns of activity. Our contribution here was to show how easy the clustering can be applied and visualized on such simple features. Developers of VLEs could use this approach to group different users by their patterns of activity, which could be applied for a variety of purposes. Bellhäuser et al. showed that personality traits like diligence and extroversion affect satisfaction and student performance [Be17]. Grouping students based on their daily activity would be a data economic way of achieving similar goals. This would allow the same results as Bellhäuser et al. without explicitly collecting and saving private attributes like psychological scales. Extroversion could e.g. be inferred from lack of activity on weekends, diligence could be connected to long streaks of activity, whereas lack of diligence could be connected to activity right before the deadline. Further research could investigate the connection between personality traits and patterns of activity.

5.1 Limitations

The primary limitation is the specific scope of the dataset. The Open University is a special distance learning university, where daily interactions with the virtual learning environment play an important role. It remains open how predictive daily activity is for settings where the learning system is not as integral. In addition to that, we focused on a simple binary classification task. Further research has to show how well the prediction tasks extents to multiple classes like "Withdrawn" and "With Distinction" and how predictive binarized daily activity is of performance indicators like grades. We also did not perform any hyperparameter optimization, which means that the scores of the machine learning systems should be regarded as lower bounds of the performance rather than optimal results.

6 Conclusion

We showed that daily activity in a VLE, even in a data economic binary representation, is a useful feature for the two machine learning tasks of predicting student's success and forming groups of users based on their learning activity. Our findings show that for these particular tasks, a healthy trade-off between big data and the principle of data minimization is possible since the binarized representation of daily activity entails sufficient information for these tasks. Future work can explore this trade-off in other contexts in higher education institutions as well as K-12. To address ethical issues like informed consent, privacy and the de-identification of data, a holistic approach that balances the potentials and perils of big data is needed.

References

[Be17] Bellhäuser, Henrik; Konert, Johannes; Röpke, René; Rensing, Christoph: Eine extravertierte und eine gewissenhafte Person in jeder Lerngruppe! Effekte der Verteilung von Persönlichkeitsmerkmalen auf Zufriedenheit und Lernergebnis. In (Igel, Christoph; Ullrich, Carsten; Martin, Wessner, eds): Bildungsräume 2017. Gesellschaft für Informatik, Bonn, pp. 309–320, 2017.

[BH18] Breiter, Andreas; Hepp, Andreas: The Complexity of Datafication: Putting Digital Traces in Context. In (Hepp, Andreas; Breiter, Andreas; Hasebrink, Uwe, eds): Communicative Figurations: Transforming Communications in Times of Deep Mediatization. Springer International Publishing, Cham, pp. 387–405, 2018.

[Ch12] Chatti, Mohamed Amine; Dyckhoff, Anna Lea; Schroeder, Ulrik; Thüs, Hendrik: A Reference Model for Learning Analytics. Int. J. Technol. Enhanc. Learn., 4(5/6):318–331, January 2012.

[D'01] D'Amico, Joseph J: A Closer Look at the Minority Achievement Gap. ERS spectrum, 19(2):4–10, 2001.

[Eu16] European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L119:1–88, May 2016.

[Fu12] Fulton, Kathleen: Upside down and inside out: Flip your classroom to improve student learning. Learning & Leading with Technology, 39(8):12–17, 2012.

[HBF18] Hepp, Andreas; Breiter, Andreas; Friemel, Thomas: Digital Traces in Context| Digital Traces in Context — An Introduction. International Journal of Communication, 12(0), 2018.

[He15] He, Jiazhen; Bailey, James; Rubinstein, Benjamin I. P.; Zhang, Rui: Identifying At-risk Students in Massive Open Online Courses. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI'15. AAAI Press, pp. 1749–1755, 2015.

[HHG15] Henrie, Curtis R; Halverson, Lisa R; Graham, Charles R: Measuring student engagement in technology-mediated learning: A review. Computers & Education, 90:36–53, 2015.

[HZZ17] Hlosta, Martin; Zdrahal, Zdenek; Zendulka, Jaroslav: Ouroboros: Early Identification of At-risk Students Without Models Based on Legacy Data. In: Proceedings of the Seventh International Learning Analytics &#38; Knowledge Conference. LAK '17, ACM, New York, NY, USA, pp. 6–15, 2017.

[Ke15] Kennedy, Gregor; Coffrin, Carleton; de Barba, Paula; Corrin, Linda: Predicting Success: How Learners' Prior Knowledge, Skills and Activities Predict MOOC Performance. In: Proceedings of the Fifth International Conference on Learning Analytics And Knowledge. LAK '15, ACM, New York, NY, USA, pp. 136–140, 2015.

[KHZ17] Kuzilek, Jakub; Hlosta, Martin; Zdrahal, Zdenek: Open University Learning Analytics dataset. Scientific Data, 4:170171, nov 2017.

[Ko13] Konert, Johannes; Burlak, Dmitrij; Göbel, Stefan; Steinmetz, Ralf: GroupAL: ein Algorithmus zur Formation und Qualitätsbewertung von Lerngruppen in E-Learning-Szenarien mittels n-dimensionaler Gütekriterien. In (Breiter, Andreas; Rensing, Christoph, eds): DeLFI 2013: Die 11 e-Learning Fachtagung Informatik. Gesellschaft für Informatik e.V., Bonn, pp. 71–82, 2013.

[KPS13] Kizilcec, René F; Piech, Chris; Schneider, Emily: Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: Proceedings of the third international conference on learning analytics and knowledge. ACM, pp. 170–179, 2013.

[KPSM17] Kizilcec, René F.; Pérez-Sanagustín, Mar; Maldonado, Jorge J.: Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. Computers Education, 104:18 – 33, 2017.

[Ku15] Kuzilek, Jakub; Hlosta, Martin; Herrmannova, Drahomira; Zdrahal, Zdenek; Wolff, Annika: OU Analyse: analysing at-risk students at The Open University. Learning Analytics Review, LAK15-1:1–16, March 2015.

[LCS] Lukarov, Vlatko; Chatti, Mohamed Amine; Schroeder, Ulrik: Learning Analytics Evaluation-Beyond Usability. DeLFI WOrkshops, pp. 123–131.

[MCM13] Michalski, Ryszard S; Carbonell, Jaime G; Mitchell, Tom M: Machine learning: An artificial intelligence approach. Springer Science & Business Media, 2013.

[Mu17] Mullis, Ina VS; Martin, Michael O; Foy, Pierre; Hooper, Martin: ePIRLS 2016: International Results in Online Informational Reading. ERIC, 2017.

[OE] OECD: PISA 2015 Results (Volume V).

[Pe11] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[PE14] Papamitsiou, Zacharoula; Economides, Anastasios A: Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society, 17(4):49, 2014.

[SP13] Slade, Sharon; Prinsloo, Paul: Learning analytics: Ethical issues and dilemmas. American Behavioral Scientist, 57(10):1510–1529, 2013.

[Wo13] Wolff, Annika; Zdrahal, Zdenek; Nikolov, Andriy; Pantucek, Michal: 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. ACM, pp. 145–149, 2013.

[Wo14a] Wolff, Annika; Zdrahal, Zdenek; Herrmannova, Drahomira; Knoth, Petr: Predicting student performance from combined data sources. In: Educational data mining, pp. 175–202. Springer, 2014.

[Wo14b] Wolff, Annika; Zdrahal, Zdenek; Herrmannova, Drahomira; Kuzilek, Jakub; Hlosta, Martin: Developing predictive models for early detection of at-risk students on distance learning modules. In: Machine Learning and Learning Analytics Workshop at The 4th International Conference on Learning Analytics and Knowledge (LAK14). 2014.