International Journal of Education and Development using Information and Communication Technology (IJEDICT), 2023, Vol. 19, Issue 2, pp. 177-190

Pragmatic monitoring of learning recession using log data of learning management systems and machine learning techniques

Khamisi Kalegele
Open University of Tanzania, Tanzania

ABSTRACT

Pragmatically, machine learning techniques can improve educators' capacity to monitor students' learning progress when applied to quality data. For developing countries, the major obstacle has been the unavailability of quality data that fits the purpose. This is partly because the in-use information systems are either not properly managed or not accessible due to ethical issues and bureaucratic processes and procedures. System log data, which is readily available for educators who use electronic learning management systems, provides opportunities for applying machine learning techniques and devising practical solutions. This study analyzed real system log data as a viable alternative for potential use in machine learning based monitoring of students who are receding in their learning. The analysis proposed several indicators of recession, and how they can be combined to ease monitoring and visualization. The proposed indicators and visualization can be used by educators to monitor students and intervene proactively.

Keywords: *Information systems; developing countries; information management; online learning*

INTRODUCTION

The COVID-19 pandemic has drastically changed the educational landscape by forcing educators to use technology to virtually provide access to learning opportunities. As Harper (2021) reports, more educators have said that technology should be a core part of their businesses. Increasingly, higher learning institutions are embracing the use of learning management systems such as Moodle (Ferdianto & Dwiniasih 2019). This kind of technology is enabling educators to provide Open and Distance Learning (ODL) opportunities to people who cannot attend traditional face to face learning institutions. The wider adoption of technologies means that more data is increasingly collected to enable applications of Artificial Intelligence (AI) in addressing chronic challenges in the education sector. As learners increase disproportionately with investments in education, many AI applications such as task automation, personalization, smart content creation, and learner support become more attractive to educators.

Unlike conventional educational institutions, ODL institutions have the advantage emanating from the necessary use of digital learning management systems. One of the advantages is the availability of various digital data that constitute profiles of learners and their digital dossier or footprints (Wigmore 2014). However, many ODL institutions have not been able to fully integrate their systems for achieving coherent digital footprints of their learners and subsequently relating them to other aspects of learning management. This means that the depth, breadth, and quality of footprint data from these systems are not known. Typically, inherent problems related to duplication, indexing, and missing data are the norm. In this article, log data from one of the best performing ODL institutions in developing countries is analyzed for the potential application of machine learning. A case study of predicting learning recession (Kalegele 2021) is used because not only does it underpin the desired outcome of

the educational institution, but it is probably the most usable application of the currently available data. This article uncovers the breadth, depth, and quality of log data from the learning management system of The Open University of Tanzania from the perspective of machine learning application.

What limits application of Machine Learning in education

Machine learning is a core part of AI that focuses on learning algorithms that use data to establish patterns that are later generalized to new or unseen data. When using the algorithms in supervised mode, at least three datasets are involved; training, validation, and test data. A training dataset is a set of examples that have been recorded and are used to fit the parameters of a machine learning model (James et al., 2013) which is denoted as $D = \{(\boldsymbol{x}_i, y_i)|i = 1{:}N\}$ where $\boldsymbol{x}$ *is a vector of training examples comprising of* M *features or attributes* and y_i *is a label of an example*. The learning process aims at fitting a machine learning model which is then evaluated in an unbiased way using a validation dataset while optimizing associated parameters (Brownlee 2020). The unbiased evaluation of the final fitted model uses a test dataset which is sometimes used interchangeably with the validation dataset.

In the past, when assessing algorithms, discussions often focused on how complex algorithms work. However, the type and quality of data used by the algorithms are equally important. There exist large living collections of databases that are used by AI and machine learning communities for empirical analyses of algorithms such as *The UCI Machine Learning Repository* (Dua & Graff 2017). Such kinds of databases have enabled discovery, benchmarking, and improvements of algorithms in profound ways. One of the most widely used datasets is IRIS (Monahan 2020). Recently, academic researchers have increased attention to the quality of data used in AI and machine learning (Richardson, Schultz, & Crawford 2019). Algorithms can only be as good as the data they use based on the garbage-in garbage-out principle. When assessing the quality of data, many issues are looked at: for example issues of completeness, accuracy, consistency, timeliness, duplication, validity, availability, and provenance (Cichy & Rass 2019). In AI and machine learning applications, these issues are broadened in scope to ensure that data is properly aggregated for learning about patterns in data, automating processes, and supporting decision-making. Moreover, data must constitute enough observations to avoid over-fitting during training (Ying 2019). The task of establishing whether data is fit-for-purpose can be daunting, and it poses an attractive challenge to researchers. In the education sector, which is the focus of this article, researchers have constantly tested various datasets to develop usable machine-learning models that solve task automation challenges.

Table 1 summarizes examples of attempts to use different data in machine learning applications in education. Moreover, Hernández-Blanco et al., (2019) compiled datasets from the year 2015 to 2018 that have been used to demonstrate how machine learning techniques can be used to improve efficiency in the education sector. Although the potential of AI in education is widely acknowledged, there are still data-related obstacles to overcome to unleash them. As noted by UNESCO (2021), any application of AI in educational contexts should properly address data related issues specific to education. However, there are also inherent issues that are likely to delay the application of AI in the education sectors of developing countries. Two issues that motivated this research are the inadequacy of data and of ethics in the education sector. Most developing countries have not been successful in managing the quality of data and its archival in digital forms. At best, educational institutions are maintaining admission and students' academic records which are not easily accessible due to ethical and privacy limitations.

***Table 1**: Previous efforts to use different data in machine learning application*

Source	Type of data	Description
Lee and Chung (2019)	School administration data	Data from central Database in Korea. It included attributes such as unauthorized absence, and unauthorized lateness.
AlZu'bi et al. (2022)	Facial images	Image classification and deep learning techniques were applied to detect the emotions of student and teachers
ONAN (2021)	Learner reviews	Conducted sentiment analysis on Massive open online courses with focus on efficiency of ensemble and deep learning
Tan and Shao (2015)	Admission Data	Statistics about enrolled and graduated students in China.
Sivakumar et al (2016)	Survey data	A dataset of 240 samples collected randomly through survey at university located at India.
Herodotou et al. (2017)	Academic performance data	Student data and their academic records
Bhardwaj et al. (2021)	Facial images	Developed algorithm for real time detection of emotions such as anger, disgust, fear, happiness, sadness, and surprise.

Learning Recession

Learning recession is a term used to refer to a condition when there is a significant decline in the learning activities of a student for an extended period (Kalegele 2021). The term has been borrowed from the field of economics in efforts to broaden the concepts of dropout and completion rates. Learning recession is a generic concept whose definition can be adapted to varying contexts such as online activities, discipline, and academic performance. In this article, the scope is learning recession in the context of online activities. Conceptually, learning recession represents worrying and undesirable episodes within the trajectory of any student's experience, including those who end up completing their programmes of study.

In ODL, online activities are the only way to measure engagement. According to many scholars, engagement is as important as grades (Estrada et al., 2021; Paulsen & McCormick 2020). When students are not engaged, grades do not matter as much because there is less assurance that they can continue learning. Over the decades, many theories that relate to students' learning experiences and achievements have been developed. Examples include student departure theory (Berger & Braxton 1998), quality of student effort (Pace 1982), student involvement theory (Astin 1984), and causal model of college environmental effects (Pascarella 1984).

Simplistically, the problem of predicting learning recession comes down to the selection of a proper indicator and differentiating recession from holidays and periods of inactivity as per the almanac of the institution. For the sake of argument, suppose the cumulative number of students is an indicator of choice, *Figure 1* and *Figure 2* show the probable impact of discovering and intervening potential recession. Using a cumulative number of online sessions, *Figure 1* shows a conceptual trend of a student. In the figure it is shown that at least students are expected to have completed 203 sessions before completion (i.e., minimum required

sessions). The figure also shows three potential episodes of learning recession, as indicated by the arrows. *Figure 2* shows a potential situation if the middle recession episode is intervened whereby the student is seen to have completed early (i.e., the student achieves 203 sessions in week 34 instead of week 43 which is 9 weeks earlier).

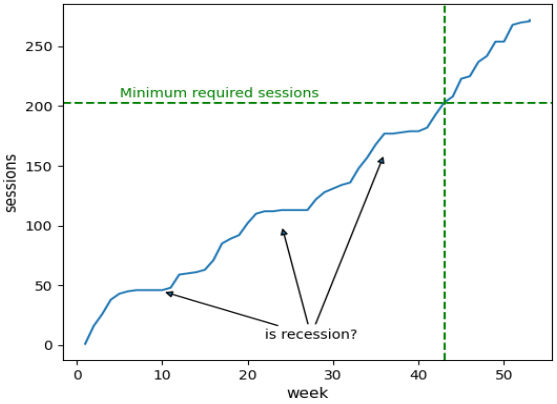

Figure 1: Conceptual cumulative sessions with probable recession episodes

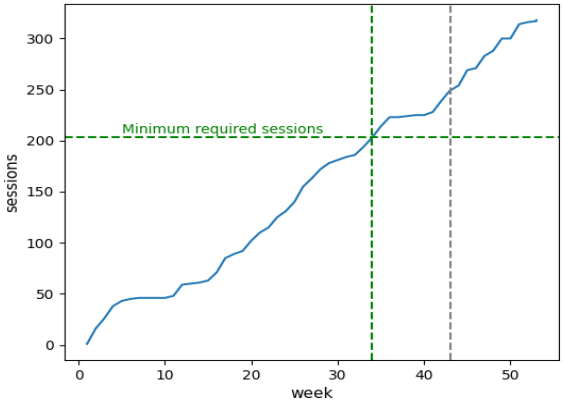

Figure 2: Conceptual cumulative sessions after overcoming recession

METHODS

The data mining approach, as explained by Shearer (2000) and Schröer, Kruse, & Gómez (2021), was used to investigate how value can be added to the digital footprints of open and distance education students. After understanding underlying learning processes, data analytics techniques were applied to digital footprint data to extract interesting patterns.

Data

The source of data is the Learning Management System (LMS) based on Moodle which has been used by the Open University of Tanzania (OUT) for more than 14 years. The LMS keeps events log data for all student's online activities. At OUT, online activities on LMS are limited to access to learning materials, submission of assignments, timed tests and quizzes, and discussion forums. Events log data consists of student identification, event timestamp, event type, and event description as shown in *Table 2*.

Table 2: *Example entry in events log*

Field	Content
Time	10/02/18, 21:51
User full name	**Surname Firstname**
Affected user	-
Event context	File: NOTES 1:Module 1
Component	File
Event name	Course module viewed
Description	.. user **53** ... viewed .. id **'943'**... course ... id **'245'**.
Origin	web
IP address	**130.24.11.26**

Bolded text is dummy data for purposes of illustration

Sampling

Events log data of students who enrolled and registered in 2015/16 were extracted from the system. A total of 1431 logs were extracted using the convenience sampling technique. The technique was preferred because of two reasons:

1. Cost and time effectiveness: Processing event log data in raw form is a time consuming and tedious process. The log data had never been analyzed before; it was not clear what to expect. Therefore, it was deemed logical to only use logs for one year of enrollment.
2. Monitoring requirements: In the interest of the study, it was necessary to include students who have been studying long enough to know their completion or graduation status.

Analysis and preparation of datasets

Mainly, analysis was done using Python pandas (https://doi.org/10.5281/zenodo.3509134 2020), its data structures, and operations. Pandas was preferred because it fitted the purpose and is an easy to use open source option. The following analysis was done.

i. Events log data of all students were combined and statistically described. For each student, the latest completion status was also established using a manual process.
ii. Key data attributes of interest were extracted from the event description. As shown in *Table 2*, they include student identification and course identification.
iii. For each student, *logged-in* and *logged-out* events were correlated to create sessions, and associated statistics were calculated.
iv. Data was cleansed by removing outliers.

Dataset for predicting completion

A standard supervised machine learning process was used to develop a prediction model. It involved four stages as follows:

1. Learning examples in the dataset were labeled as *graduated* or *continuing* based on the information obtained from the University authority.

2. 10% of the dataset was set aside as a test set while the remaining 90% was divided into training and validation sets in an 8:2 ratio respectively.

3. Models were developed using Logistic regression, K-Nearest Neighbors, Support Vector Machine (SVM) classifier, Naive Bayes, Multilayer Perceptron (MLP) neural network, and Random Forest algorithms. Although this article is not investigating the performance of algorithms, these were selected because the literature indicated that they have been used in education. Scikit-Learn API (Pedregosa et al., 2011) was used to provide implementations of the specific algorithms.

4. During training, hyperparameters were optimized using a randomized search cross-validation approach. The approach allows for the definition of a search space as a bounded domain of hyperparameter values which are then randomly selected. Scikit-Learn API was used to provide the implementation of the randomized search cross-validation.

5. *k* -fold cross-validation (Allen 1974) was used to estimate skill of each of the developed models on unseen data. The value of *k* was set to 10 because it is the most used and its performance has recently been validated (Marcot & Hanea 2021).

Recession detection

Detection of recession in a student's learning is a complicated task because it is often rare and therefore must be dealt with from unbalanced data. Nonetheless, it is important to be done because it can inform regulations of learning institutions in efforts to enact countermeasures. As a starting point, this study formulated a simple but pragmatic solution for a complex problem, following in the footsteps of many scholars who believe that the best solution is one which is built from simplicity. Thus, this study assumes that a lapse in student engagement indicates a recession in some way. The study opts for an unsupervised learning approach instead of a supervised one to avoid the complications of dealing with unbalanced data for rare events. According to Mahesh (2020), two commonly used unsupervised approaches are clustering (e.g., K-Means) and dimensionality reduction such as Principal Component Analysis (PCA). In this work, PCA based algorithm was used to find a projection of the data that maximizes the variance and obtain a new single variable called recession delta. The projection axis is then used to project any new data which will be seen whereby the magnitude of the recession delta will indicate the level of engagement.

RESULTS AND DISCUSSION

Online sessions

Digital footprint data for 1431 students were extracted from the logs of the learning management system. The data included user-id, session start time, session end time, session duration, and session description. Example data items are shown in *Table 3* where tags (e.g., list 180 user-id1) were used to represent session descriptions as lists of associated events. From the data, it was established that there are an average of 218 active students in a month, each of whom engages with the system 4 times.

Table 3: *Example details of sessions*

session id	start time	end time	duration	description
180 userid1	15/03/2021 13:44	15/03/2021 14:47	63	list 180 userid1
179 userid1	14/03/2021 15:03	14/03/2021 15:09	6	list 179 userid1
178 userid2	12/03/2021 16:39	12/03/2021 18:08	89	list 178 userid2

Features selection

From the sessions, 9 attributes were selected: number of sessions (NS), mean session duration (ASD), total study time (ST), total course views (NCV), unique course views (UCV), number of assignments (NA), number of submissions (NAS), forum activities (NFA), and public private IP ratio (PPR). The correlation matrix of these attributes is shown in *Figure 3* below.

Data aggregation options

Timestamping of the log data offers several aggregation options giving users views of interest into the data. After encoding the data with one-hot encoding for categorical features and scaling numerical features using StandardScaler, four aggregation options were confirmed to be plausible. These are shown graphically in *Figure 4* below. For this study, option 4 was used whereby aggregation was done monthly per programme.

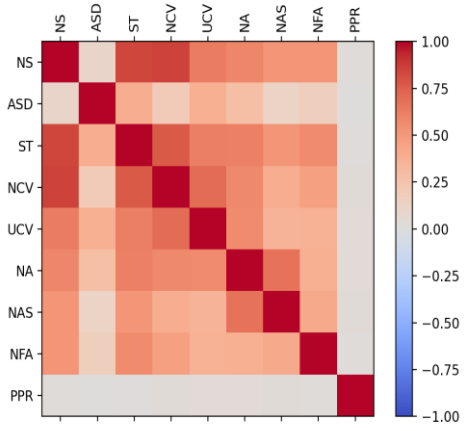

***Figure 3*:** Correlation of attributes

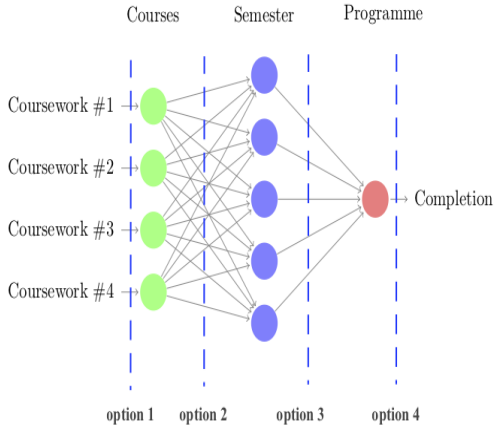

Figure 4: Data aggregation options

Recession delta

For detecting recession in learning, six (6) additional features were included in the analytics: number of active months (NAM), number of inactive months (NIM), log timespan in days (LD), gender (G), and programme (P). This followed the observation that there are long inactive periods within the digital footprints of the students. Moreover, gender influences the dynamics of online engagements as shown in *Figure 6* below. In the figure, it can be noted that more females abscond than males while more males postpone studies more than females. Also, the proportion of females who graduate is higher than the proportion of females who continue.

Equation 1:

$$\Delta d_i = d_i - \bar{d} \; where \; 0 < i < N \qquad 1$$

Equation 2:

$$\text{Level of engagement, } e_i = \begin{cases} \text{Significantly more engaged if } \Delta d_i \geq 0.75 \\ \text{Slightly more engaged if } 0 \leq \Delta d_i < 0.75 \\ \text{Slightly less engaged if } -0.75 < \Delta d_i < 0 \\ \text{Significantly less engaged if } -0.75 \leq \Delta d_i \end{cases} \qquad 2$$

The PCA algorithm based projection model had reconstruction error (mean squared distance) of 5.09 on training dataset and 4.80 on the test dataset. Using the projection, for each programme of study with total number of students N , and for a certain calendar month, the average one dimension value was calculated as the *programme mean dimension*, $\bar{d}$. Students' reduced dimension values for the month was compared to the programme's overall value to obtain their recession delta Δd as shown by Equation 1 above. Normalized values of delta were then used to determine level of engagement using a simple threshold values as shown in Equation 2 above using 25% (any threshold value can be used depending on the need of the institutions). This allowed the study to categorize students in a monthly basis for four years based on their digital footprint data. Example categorization using a programme of Bachelor of Arts in Public Administration is shown in *Figure 5*.

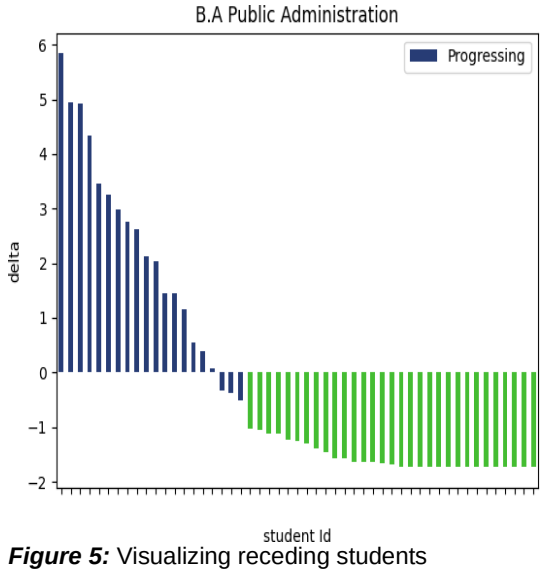

Figure 5: Visualizing receding students

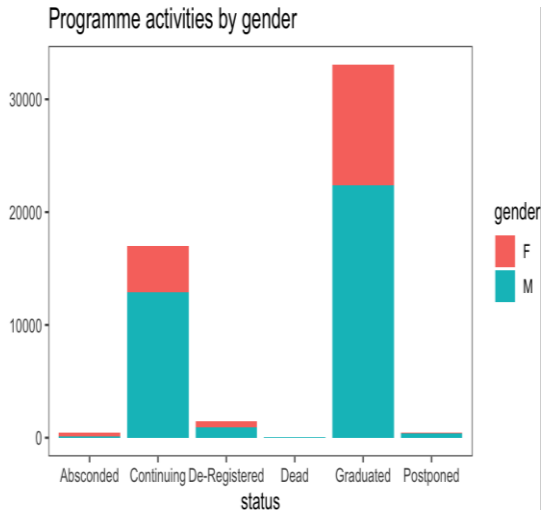

Figure 6: Volume of online activities by gender

Completion prediction model

Using the processed data, the study compared several learning algorithms for pragmatic use in predicting how many students were likely to graduate. The study found that, in terms of F1 Score, SVM and MultiLayer Perceptron (MLP) performed better than other algorithms (i.e., Random Forest, Logistic Regression, KNN, Gradient boosting, and Naive Bayes). Training performance also exhibited a similar order except for KNN and Random Forest. That is, although KNN offered less performance in terms of F1 Score, its training was slightly more efficient and less sensitive.

The data in Table 4 below summarizes the training and performance metrics for all tested algorithms.

Table 4: *Training performances of tested algorithms*

Algorithm	Training Acc	X-validation Acc	AUC	F1 score
SVM classifier	86.02	74.68	0.9068	0.8952
MLP	83.75	78.65	0.8830	0.8700
Gradient boosting	76.79	74.51	0.8434	0.8259
Random forest	75.19	73.74	0.8298	0.8231
K-NN	77.13	73.57	0.8347	0.8172
Logistic Regression	75.19	73.75	0.8024	0.8172
Naive Bayes	72.31	71.20	0.7566	0.7960

DISCUSSION

The potential for applying machine learning to improve education was manifesting quite slowly before the COVID-19 pandemic. One of the reasons was the lack of accurate and quality

data that will enable the development of practical machine learning models. The pandemic has drastically changed educational landscapes by forcing educators to use technology to virtually provide access to learning opportunities. This is resulting in the accumulation of quality digital footprint data, attracting a revisit of previous studies to validate results (Thanh Noi & Kappas 2018) and new aspects such as the detection of undesirable emotions during learning sessions (AlZu'bi et al. 2022; Bhardwaj et al. 2021).

As much as the overall aim of applying machine learning in education has remained on achieving precision education, its success also remains dependent on mining digital footprint data. In most cases, existing approaches are incomparable as they differ in breadth and depth of features as well as context. Contrary to other studies in the same context (Mwalumbwe & Mtebe 2017),the inclusion of course specific attributes has resulted in four data aggregation options. Although this study dealt with the option that focuses on analytics of study programmes viewed monthly, the approach enables analytics at course and semester levels.

Recession delta, presented in Equations 1 and 2, is an interesting attempt to change the paradigm whereby the central focus has mainly been on determining progression and completion rate. Because students at The Open University of Tanzania are mostly working adults with family, it was realized that every episode of recession is potentially detrimental to their overall progression. This is clearly explained in *Figure 1* and *Figure 2* where a recession intervention leads to early completion. Moreover, the reported high enrollment rates at open and distance learning institutions often amount to vanity metrics because experts suggest that factors such as student engagement and experience matter a lot (Jordan 2015; Hernández-Blanco et al., 2019). In this study, Equation 2 enables the proposed approach to track engagement levels for early intervention.

Algorithms that were tested in this study yielded positive results for pragmatic applications. Validation performances were consistent with training performances except for KNN and Random Forest whereas KNN offered less performance in terms of F1 Score while its training was slightly more efficient and less sensitive. This observation validated previous studies (Thanh Noi & Kappas 2018) which investigated performances of non-parametric algorithms and noted that the two algorithms offer mixed levels of performance. For pragmatic applications using digital footprint data from learning management systems, most of the non-parametric and supervised algorithms can be used and will offer varying performances depending on levels of optimization and quality of data

CONCLUSION

Key contributions

This paper presents a viable alternative, which is relatively cheap, for tracking students' learning and enabling instructors to identify students who need interventions. If systems are properly configured to log events correctly, log data can be harnessed to facilitate some learning management activities. It validates observations by other researchers that learning analytics using digital footprints can help to identify struggling students much earlier. For instance, Purdue University in the United States was using learning analytics to identify problems as early as the second week by merging information known about individuals and their digital footprint. (Sclater, Peasgood & Mullan 2016; Pistilli & Arnold 2010). In our case, the challenge of proper data management is preventing easy access to student information. Resources are needed to curate most data before they can be used, something which will take time. Among the key findings:

- This study observed that the proportion of females who graduate is higher than the proportion of females who continue. An additional study is needed to correlate this observation with new information from students.
- This paper proposes a set of attributes and associated processing which can be aggregated at four different points (i.e., coursework level, courses level, semester, and programme level as shown in *Figure 4*) to generate a variety of datasets for further application of analytics. A previous study in a similar context of developing countries was limited to few statistics which included *login frequency*, *time spent*, *number of downloads*, *interactions with peers*, *number of performed exercise*, and *number of forum posts* (Mwalumbwe & Mtebe 2017). This study introduced various new features for the context as explained in Sections 2 and 3. Potentially, applying proper analytics on data can be used to develop various management dashboards.
- This paper also proposes the use of Principal Component Analysis on log data to reduce its dimensionality and enable visualization of learning recession as explained in previous subsections.
- This paper presents a handy technique for processing log data from a Moodle-based learning management system hinged on session statistics and a corpus or list of events as shown in *Table 3*.
- This paper depicts that log data alone, without details of students' grades, cannot produce completion prediction models with the desired level of performance for any practical use.

Limitations

- Prediction of students' completion of study and detection of learning recession that are presented in this paper are limited to the use of log data only. The study focused on understanding the breadth and depth of the log data. If log data is supplemented with academic grades, better performances will be attained.
- The produced models were only cross validated and not tested against students in subsequent years. Open universities provide golden opportunities to test these kinds of models. However, the most benefit will be obtained if models are built from more integrated datasets as is with Open University Analyse (Kuzilek et al., 2015), and the lowest aggregation point (i.e., Option 1 in *Figure 4*).
- The graduation prediction model could be adjusted for optimal precision-recall balance to achieve better performance. This was not achieved in the current study because we realized that more data and features would be more useful, but it is something that will not be possible soon. An interesting new study will be the creation of richer datasets per specific programmes of study from real life experiences of open universities.

KEY TAKEAWAYS

This paper presents a case of using digital footprint data from the learning management system to enable interventions for learning recession and the inability to complete studies on time. In the paper, it is shown that log data constitutes sufficient digital footprint to allow such use as detecting learning recession. The approach to detect recession based on Principal Component Analysis which was established pragmatically was shown to be viable for practical use. It is shown that not enough performance is achieved when the data is used to predict

the prospects of completion of studies.The research also generated useful datasets which promote further discussion and research.

REFERENCES

Allen, David M. (1974). "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction". In: *Technometrics* vol. 16, no.1, pp. 125–127. DOI: 10.1080/00401706. 1974.10489157.

AlZu'bi, S.; Abu Zitar, R.; Hawashin, B.; Abu Shanab, S.; Zraiqat, A.; Mughaid, A.; Almotairi, K.H.; Abualigah, L. (2022). "A Novel Deep Learning Technique for Detecting Emotional Impact in Online Education". *Electronics*, vol. 11, no. 18, pp. 2964. https://doi.org/10.3390/electronics11182964

Astin, Alexander (Jan. 1984). "Student Involvement: A Development Theory for Higher Education". In: *Journal of College Student Development*, vol. 40, pp. 518–529.

Berger, J. B., & Braxton, J.M. (1998). "Revising Tinto's Interactionalist Theory of Student Departure through Theory Elaboration: Examining the Role of Organizational Attributes in the Persistence Process". In: *Research in Higher Education* vol. 39, no. 2, pp. 103–119. ISSN: 03610365, 1573188X. URL: http://www.jstor.org/stable/40196289.

Bhardwaj, P., Gupta, P.K., Panwar, H., Siddiqui, M.K., Morales-Menendez & Bhaik, A. (2021). "Application of Deep Learning on Student Engagement in e- learning environments". In: *Computers and Electrical Engineering* vol. 93, p. 107277. ISSN: 0045- 7906. DOI: https://doi.org/10.1016/j.compeleceng.2021.107277. URL: https://www.sciencedirect.com/science/article/pii/S0045790621002597.

Brownlee, J. (Aug. 2020). What is the Difference Between Test and Validation Datasets? URL: https://machinelearningmastery.com/difference-test-validation-datasets/.

Cichy, C., & Rass, S. (2019). "An Overview of Data Quality Frameworks". In: *IEEE Access*, vol. 7, pp. 24634–24648. DOI: 10.1109/ACCESS.2019.2899751.

Dua, D., & Graff, C. (2017). UCI Machine Learning Repository. URL: http://archive.ics.uci.edu/ml

Estrada, M.; Monferrer, D.; Rodríguez, A.; Moliner, M.Á. (2021). "Does Emotional Intelligence Influence Academic Performance? The Role of Compassion and Engagement in Education for Sustainable Development". In: *Sustainability*, vol. 13, no. 4, ISSN:2071-1050. UR: https://www.mdpi.com/2071-1050/13/4/1721.

Ferdianto, F., & Dwiniasih (Oct. 2019). "Learning Management System (LMS) schoology: Why it's important and what it looks like". In: *Journal of Physics: Conference Series* 1360.1, p. 012034. doi: 10.1088/1742-6596/1360/1/012034. url: https://doi.org/10.1088/1742-6596/1360/ 1/012034.

Harper, T. (Aug. 2021). Top 7 Ways Artificial Intelligence Is Used in Education. url: https://trainingmag.com/top-7-ways-artificial-intelligence-is-used-in- education/.

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). "A Systematic Review of Deep Learning Approaches to Educational Data Mining". In: Complexity 2019, p. 1306039. doi: 10.1155/2019/1306039. url: https://doi.org/10.1155/2019/1306039.

Herodotou, C., Gilmour, A., Boroowa, A., Rienties, B., Zdrahal, Z., & Hlosta, M.(2017). "Predictive modelling for addressing students attritionin Higher Education: The case of OU Analyse". In: CALRG Annual Conference 2017. URL: http://oro.open.ac.uk/49470/

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer. URL: https://faculty.marshall.usc.edu/gareth-james/ISL/

Jordan, K. (June 2015). "Massive open online course completion rates revisited: Assessment,length and attrition". In: *The International Review of Research in Open and Distributed Learning*, vol. 16, no. 3, DOI: 10.19173/irrodl.v16i3.2112, URL: https://www.irrodl.org/index.php/irrodl/article/view/2112.

Kalegele, K. (2021). "Analysis of ODL Logs for Predicting Recession". In: 2021 IEEE AFRICON,Arusha, Tanzania, United Republic of, September 13-15, 2021. IEEE, pp. 1–4. DOI: 10.1109/ AFRICON51333.2021.9571018. URL: https://doi.org/10.1109/AFRICON51333.2021.9571018.

Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrahal, Z., Vaclavek, J., & Wolff, A. (2015). "OU Analyse: analysing at-risk students at The Open University". In: Learning Analytics Review LAK15-1. Presented on Wednesday 18th March 2015 in the Students At Risk session and on Thursday 19th March 2015 in the Technology Showcase session of the 5th International Learning Analytics and Knowledge (LAK) Conference:Scaling Up: Big Data to Big Impact, held at Poughkeepsie, New York (USA), pp. 1–16. URL: http://oro.open.ac.uk/42529/.

Lee, S., & Jae Young Chung (July 2019). "The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction". In: *Applied Sciences* vol. 9, p. 3093. DOI: 10.3390/app9153093.

Mahesh, B. (Jan. 2020). "Machine Learning Algorithms - A Review". In: *Computational Statistics*, vol. 9, pp. 381–386. DOI: 10.1007/s00180-020-00999-9.

Marcot, B., & Hanea, A. (2021). "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?" In: *International Journal of Science and Research* (IJSR), vol. 36.

Monahan, Kyle M. (2020). Iris dataset for machine learning. Version V1. doi: 10.7910/DVN/R2RGXR. url: https://doi.org/10.7910/DVN/R2RGXR.

Mwalumbwe, I., & Mtebe, J. (Mar. 2017). "Using Learning Analytics to Predict Students' Performance in Moodle Learning Management System: A Case of Mbeya University of Science and Technology". In: *Electronic Journal of Information Systems in Developing Countries*, vol. 79, pp. 1–13. DOI: 10.1002/j.1681-4835.2017.tb00577.x.

ONAN, A. (2021). "Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach". In: *Computer Applications in Engineering Education* 29.3, pp. 572–589. doi: https://doi.org/10.1002/cae.22253. eprint: https://onlinelibrary.wiley. com/doi/pdf/10.1002/cae.22253. url: https://onlinelibrary.wiley.com/doi/abs/10. 1002/cae.22253.

Pace, C. Robert and DC. National Commission on Excellence in Education (ED) Washington (1982). Achievement and the Quality of Student Effort [microform] / C. Robert Pace. English. Distributed by ERIC Clearinghouse [Washington, D.C.], 40 p. url: https://eric.ed.gov/?id= ED227101.

Pascarella, E., T. (1984). "College Environmental Influences on Students' Educational Aspirations". In: *The Journal of Higher Education*, vol. 55, no. 6, pp. 751–771. ISSN: 00221546, 15384640. URL: http://www.jstor.org/stable/1981512.

Paulsen, J., & McCormick, A.C. (2020). "Reassessing Disparities in Online Learner Student Engagement in Higher Education". In: *Educational Researcher* vol. 49, no. 1, pp. 20–29. DOI: 10.3102/0013189X19898690. eprint: https://doi.org/10.3102/0013189X19898690.

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.(2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* vol. 12, pp. 2825–2830.

.

Pistilli, Matthew D., &. Arnold, Kimberly E. (2010). "Purdue Signals: Mining Real-Time Academic Data to Enhance Student Success". In: *About Campus*, vol. 15, No. 3, pp. 22–24. DOI: 10.1002/abc.20025. eprint: https://doi.org/10.1002/abc.20025.

Richardson, R., J. Schultz, and Crawford, K. (2019). "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice". In: 94 N.Y.U. L. REV. ONLINE 192.

Schröer, C., Felix Kruse, F., & Gomez, J.M.(2021). "A Systematic Literature Review on Applying CRISP-DM Process Model". In: Procedia Computer Science 181. CENTERIS 2020 - International Conference on Enterprise Information Systems / ProjMAN 2020 - International Conference on Project Management / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020, pp. 526–534. issn: 1877-0509. doi: https://doi.org/10.1016/j.procs.2021.01.199. url: https://www.sciencedirect.com/science/article/pii/S1877050921002416.

Sclater, N., Peasgood, A., & Mullan, J. (2016). Learning analytics in higher education: a review of UK and international practice. Tech. rep. Bristol, 39 p. URL:https://www.jisc.ac

Shearer, Colin (2000). "The CRISP-DM Model: The New Blueprint for Data Mining". In: *Journalof Data Warehousing* vol. 5, no. 4

Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). "Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree". In: *Indian Journal of Science and Technology* vol. 9, pp. 1–5. doi: 10.17485/ijst/2016/ v9i4/87032.

Tan, M., & Shao, P. (2015). "Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method". In: *International Journal of Emerging Technologies in Learning* (iJET) 10.1, pp. 11–17. ISSN: 1863-0383. url: https://online-journals.org/index. php/i-jet/article/view/4189.

Thanh Noi, P., & Kappas, M. (2018). "Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery". In: Sensors 18.1. issn: 1424-8220. doi: 10.3390/s18010018. url: https://www.mdpi.com/ 1424-8220/18/1/18.

The pandas development (Feb. 2020). pandas-dev/pandas: Pandas. Version latest. doi: 10.5281/zenodo.3509134. url: https://doi.org/10.5281/zenodo.3509134.

UNESCO (2021). AI and education: guidance for policymakers. UNESCO. url: https://unesdoc. unesco.org/ark:/48223/pf0000376709.

Wigmore, I. (May 2014). What is digital footprint - definition from whatis.com. url: https ://www.techtarget.com/whatis/definition/digital-footprint.

Ying, Xue (Feb. 2019). "An Overview of Overfitting and its Solutions". In: *Journal of Physics*: Conference Series 1168, p. 022022. doi: 10.1088/1742-6596/1168/2/022022. url: https://doi.org/10.1088/1742-6596/1168/2/022022.