DATA SCIENCE JOURNAL

Mduma, N, et al. 2019. A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction. *Data Science Journal,* 18: 14, pp. 1–10. DOI: https://doi.org/10.5334/dsj-2019-014

REVIEW

A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction

Neema Mduma[1], Khamisi Kalegele[2] and Dina Machuve[1]

[1] School of Computational and Communication Science and Engineering, NM-AIST, Arusha, TZ

[2] Tanzania Commission for Science and Technology, COSTECH, Dar es salaam, TZ

Corresponding author: Neema Mduma (mduman@nm-aist.ac.tz)

School dropout is absenteeism from school for no good reason for a continuous number of days. Addressing this challenge requires a thorough understanding of the underlying issues and effective planning for interventions. Over the years machine learning has gained much attention on addressing the problem of students dropout. This is because machine learning techniques can effectively facilitate determination of at-risk students and timely planning for interventions. In order to collect, organize, and synthesize existing knowledge in the field of machine learning on addressing student dropout; literature in academic journals, books and case studies have been surveyed. The survey reveal that, several machine learning algorithms have been proposed in literature. However, most of those algorithms have been developed and tested in developed countries. Hence, developing countries are facing lack of research on the use of machine learning on addressing this problem. Furthermore, many studies focus on addressing student dropout using student level datasets. However, developing countries need to include school level datasets due to the issue of limited resources. Therefore, this paper presents an overview of machine learning in education with the focus on techniques for student dropout prediction. Furthermore, the paper highlights open challenges for future research directions.

Keywords: Machine Learning (ML); Imbalanced learning classification; Secondary education

1 Introduction

Reducing student dropout rates is one of the challenges facing in the education sector globally. The problem has brought a major concern in the field of education and policy-making communities (Aulck et al., 2016). A growing body of literature indicates high rates of students dropout of school especially pronounced in the developing world; with higher rates for girls compared to boys in most parts of the world (Shahidul and Karim 2015). In Tanzania, for example, student dropout is higher in lower secondary education compared to higher level where girls are much less likely to finish secondary education comparing to boys; 30% of girls dropout before reaching form 4 as compared to 15% percent for boys (President's Office et al., 2016). The scenario is different in primary education, where by boys tend to drop-out of school more compared to girls. Besides, to the knowledge of searchers, developing countries lack enough researches on addressing this problem in higher level education. Finding and implementing solutions to this problem has implications well beyond the benefits to individual students. Moreover, enabling students to complete their education means investing in future progress and better standards of life with multiplier effects. To effectively address this problem, it is crucial to ensure that all students finish their school on time through early intervention on students who might be at risk of dropping classes. This require data-driven predictive techniques that can facilitate determination of at-risk students and timely planning for interventions (Fei and Yeung 2015).

Machine learning approaches are one of the well sought solutions to addressing school dropout challenge. Various studies have been conducted in developed countries on developing student predictive algorithms (Adhatrao et al., 2013; Durairaj and Vijitha, 2014; Chen et al., 2014). Moreover, there exist quite a significant body of literature on machine learning based approaches associated with fighting dropouts (Sales et al., 2016; Lakkaraju et al., 2015; Ameri et al., 2016). The knowledge embodied in literature has the potential to transform the fight against dropout from reactive to proactive. This is a more feasible now

than ever because the Information and Communication Technologies (ICTs) have already transformed the way data has been collected and managed, which is a key ingredient to any intelligent harnessing of useful patterns of recorded events. Despite several efforts done by previous researchers, there are still challenges which need to be addressed. Most of the widely used datasets are generated from developed countries. However, developing countries are facing several challenges on generating public datasets to be used on addressing this problem. Cost and time consuming are factors that led data collection process to be very difficult. The study conducted by Mgala (2016) used the primary education data collected in Kenya, although the dataset is not publicly available. Besides that, Uwezo data on learning[1] is the publicly available dataset which was collected countrywide for primary schools in Tanzania. The dataset focused on individual household data, including education.

In developing countries, prospects of dropout-free education system are still slim considering the scale of socio economic challenges, which are deemed central to the retention of students in schools. Increasingly, communities of practitioners and researchers are looking at machine learning approaches as a likely solution for achieving dropout-free schools. Student dropout has been a serious problem that adversely affects the development of the education sector, this is due to a complex interplay of socio-cultural, economic and structural factors (Mosha, 2014). Schooling, according to the human capital theory, is an investment that generates higher future income for individuals (Patron, 2014). Many developing countries are experiencing high dropout rate of secondary school students as a big challenge which has been considered as a problem for the individual and society (Halland et al., 2015). However, less attention is paid to improve quality of education to people belongs to any class. In this regard, a UNESCO (2011) report points out, that about one thirty million children in the developing world denied their right to education through dropping out (Latif et al., 2015).

In responding to this problem of dropping out and other challenges facing secondary schools, Tanzania as one among developing countries introduced an Education Training Policy (ETP) and Education Sector Development Plan (ESDP) (TAMISEMI, 2004). These were established to focus on access, quality improvement, capacity development and direct funding to secondary schools. The combined effort was expected to improve the overall status of secondary education, but still the problem is far from over.

Therefore, in this article a survey of how machine learning techniques have been used in the fight against dropouts is presented. The purpose of conducted survey is to provide a stepping-stone for students, researchers and developers who aspire to apply the techniques. Key intervention points that were identified during our preliminary survey guided the herein presented survey. The intervention points included issues related to algorithms for predicting dropouts.

2 Method of study

This paper surveys the literature in academic journals, books, and case studies. The objective is to collect, organize, and synthesize existing knowledge relating to machine learning approaches on student dropout prediction. The surveyed papers focused on several works which have been done on machine learning in education such as student dropout prediction, student academic performance prediction, student final result prediction etc. The findings of these studies are very useful on understanding the problem and improving measures to address solution. We searched several databases such as ResearchGate, Elsevier, Association for Computing Machinery (ACM), Science Direct, Springer Link, IEEE Xplore, and other computer science journals. In searching sentences and keywords we used predicting student dropout, predicting student dropout using machine learning techniques, application of machine learning in education and student dropout prediction using machine learning techniques. We examined each article's reference list to identify any potentially relevant research or journal title. The publication periods taken into consideration is 2013 to 2017. On types of text searched we use PDF, Documents and Full length paper with abstract and keywords. Furthermore, in search items we used journal articles, conferences paper, workshop papers, topics related blogs, expert lectures or talks and other topic related communities such as educational machine learning community. A substantial subset of the culled articles contributed to warrant inclusion in this study.

3 Machine learning in education

Over the past two decades, there has been significant advances in the field of machine learning. This field emerged as the method of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications (Jordan and Mitchell 2015). There are several areas where machine learning can positively impact education. The study conducted by Center

[1] http://www.twaweza.org/go/uwezo-datasets.

for Digital Technology and Management (2015), reported on the growth of the use of machine learning in education, this is due to the rise in the amount of education data available through digitization. Various schools have started to create personalized learning experiences through the use of technology in classrooms. Furthermore, Massive open on-line courses (MOOCs) have attracted millions of learners and present an opportunity to apply and develop machine learning methods towards improving student learning outcomes and leveraging the data collected (Lee 2017).

Owing to the advancement of the amount of data collected, machine learning techniques have been applied to improve educational quality including areas related to learning and content analytics (Lan et al., 2014; Waters et al., 2014), knowledge tracing (Yudelson et al., 2013), learning material enhancement (Rakesh et al., 2014) and early warning systems (Beck and Davidson 2016; Brundage, 2014; US Department of Education, 2016). The use of these techniques for educational purpose is a promising field aimed at developing methods of exploring data from computational educational settings and discovering meaningful patterns (Nunn et al., 2016).

One of the first applications of machine learning in education had been helping quizzes and tests move from multiple choice to fill in the blank answers.[2] The evaluation of students' free form answers was based on Natural Language Processing (NLP) and machine learning. Various studies on efficacy of automated scoring show better results than human graders in some cases. Furthermore, automated scoring provides more immediate scoring than a human, which helps for use in formative assessment.

A few years ago, prediction has been observed as an application of machine learning in education.[3] A research conducted by Kotsiantis (2012), presented a novel case study describing the emerging field of educational machine learning. In this study, students' key demographic characteristic data and grading data were explored as the data set for a machine learning regression method that was used to predict a student's future performance. In a similar vein, several projects were conducted including a project that aims to develop a prediction model that can be used by educators, schools, and policy makers to predict the risk of a student to drop out of school.[4] Springboarding from these examples, IBM's Chalapathy Neti shared IBM's vision of Smart Classrooms using cloud-based learning systems that can help teachers identify students who are most at risk of dropping out, and observe why they are struggling, as well as provide insight into the interventions needed to overcome their learning challenges.[5]

Certainly, machine learning application in education still face several challenges that need to be addressed. There is lack of available open-access datasets especially in developing countries; more data-sets need to be developed, however cost must be acquired. Apart from that, several researchers ignore the fact that evaluation procedures and metrics should be relevant to school administrators. According to Lakkaraju et al. (2015), the evaluation process should be designed to cater the needs of educators rather than only focused on common used machine learning metrics. In addition to that; the same study reveals that, many studies focused only on providing early prediction. While, a more robust and comprehensive early warning systems should be capable of identifying students at risk in future cohorts, rank students according to their probability of dropping and identifying students who are at risk even before they drop. Therefore, developing countries need to focus on facilitating a more robust and comprehensive early warning systems for students' dropout. Also, there is need to focus on school level datasets rather than only focusing on student level datasets; this is due to the fact that school districts often have limited resources for assisting students and the availability of these resources varies with time. Therefore, identifying at risk schools will help the authorities to plan for resource allocation before the risk.

Furthermore, in the context of education data imbalance is very common classification problem in the field of student retention, mainly because the number of registered students is large compared to the number of dropout students (Thammasiri et al., 2014). According to Gao (2015), the imbalanced ratio is about at least 1:10. Besides, the minority class usually represents the most important concept to be learned, it is difficult to identify it due to exceptional and significant cases (López et al., 2013). Since accuracy as a widely used metric has less effect on minority class than majority class (Longadge et al., 2013; Lin and Chen, 2013), several researchers applied other metrics such as F-measure (Mgala and Mbogho 2015; Rovira et al., 2017; Aulck et al., 2017), Mean Absolute Error (MAE) (Ameri et al., 2016; Elbadrawy et al., 2016; Lakkaraju et al., 2015; Rovira et al., 2017), Area Under the curve (AUC) (Liang et al., 2016; Fei and Yeung, 2015; Aulck et al., 2016; Prieto et al.,

[2] http://www.gettingsmart.com/2017/04/next-big-thing-education/.
[3] https://www.linkedin.com/pulse/ai-classroom-machine-learning-education-michael-s-davison-iii.
[4] https://2016.hackerspace.govhack.org/content/early-dropout-prediction-higher-education-using-machine-learning-approach-australian-case.
[5] http://www.research.ibm.com/cognitive-computing/machine-learning-applications/decision-support-education.shtml.

2017; Mgala and Mbogho, 2015; Halland et al., 2015), mean squared error (Iam-On and Boongoen 2017; Xu et al. 2017), Root-Mean-Square Error (RMSE) (Elbadrawy et al., 2016), error residuals (Poh and Smythe 2015), and misclassification rates (Hung et al., 2017) on addressing the problem of student dropout.

The power of machine learning can step in building better data to help authorities draw out crucial insights that change outcomes. When students drop out of school instead of continuing with education, both students and communities lose out on skills, talent and innovation.[6] On addressing student dropout problem, several predictive models were developed in developed countries to process complex data sets that include details about enrollment, student performance, gender and socio-economic demographics, school infrastructure and teacher skills to find predictive patterns. Although on developing predictive models, developing countries need to consider other factors such as school distance which has been ignored by several researchers but matters in the developing countries' scenario. Despite the fact that, evaluation of developed predictive models tend to differ but the focus remains on supporting administrators and educators to intervene and target the most at-risk students so as to invest and prevent dropouts in order to keep young people learning.

4 Machine learning techniques on addressing student dropout

In the context of education on addressing student dropout prediction, the techniques for learning can be supervised or unsupervised.

Supervised learning is based on learning from a set of labeled examples in the training set so that it can identify unlabeled examples in the test set with the highest possible accuracy (Erik G., 2014). The paradigm of this learning is efficient and it always finds solutions to several linear and non-linear problems such as classification, plant control, forecasting, prediction, robotics and so many others (Sathya and Abraham 2013).

Several existing works have focused on supervised learning algorithms such as Naive Bayesian Algorithm, Association rules mining, ANN based algorithm, Logistic Regression, CART, C4.5, J48, (BayesNet), SimpleLogistic, JRip, RandomForest, Logistic regression analysis, ICRM2 for the classification of the educational dropout student (Kumar et al., 2017). However, under the classification techniques, Neural Network and Decision Tree are the two methods highly used by the researchers for predicting students' performance (Shahiri et al., 2015; Joseph 2014). The advantage of neural network is that, it has the ability to detect all possible interactions between predictors variables (Gray et al., 2014) and could also perform a complete detection without having any doubt even in complex nonlinear relationship between dependent and independent variables (Arsad, Pauziah Mohd Buniyamin, Norlida Manan, 2013), while decision tree had been used because of its simplicity and comprehensibility to uncover small or large data structure and predict the value (Natek and Zwilling, 2014).

Unlike supervised, unsupervised learning algorithm is used to identify hidden patterns in unlabeled input data. It refers to provide ability to learn and organize information without an error signal and be able to evaluate the potential solution. The lack of direction for the learning algorithm in unsupervised learning can sometime be advantageous, since it lets the algorithm to look back for patterns that have not been previously considered (Sathya and Abraham, 2013).

Several techniques have been proposed on addressing this problem of student dropout using different approaches such as Survival Analysis (Ameri, 2015; Ameri et al., 2016), Matrix Factorization (Iam-On and Boongoen, 2017; Hu and Rangwala, 2017; Elbadrawy et al., 2016; Iqbal et al., 2017; Babu 2015), and Deep Neural Network (Fei and Yeung 2015; Wang et al. 2017b). Other approaches such as time series clustering (Hung et al., 2017; Młynarska et al., 2016) were presented to perform clustering, which are extensively used in recommender systems (Xu et al., 2017).

Survival analysis is used to analyze data in which the time until the event is of interest (Kartal 2015). It provides various mechanisms to handle such censored data problems that arise in modeling such as longitudinal data (also referred as time-to-event data when modeling a particular event of interest is the main objective of the problem) which occurs ubiquitously in various real-world application domains (Wang et al., 2017a).

In the context of education, the use of survival analysis modeling to study student retention was developed. Ameri et al. (2016) developed a survival analysis framework with the aim of identifying at-risk students using Cox proportional hazards model (Cox) and applied time-dependent Cox (TD-Cox). This approach captures time-varying factors and leverage those information to provide more accurate prediction of student dropout, using the dataset of students enrolled at Wayne State University (WSU) starting from 2002 until

[6] https://www.microsoft.com/empowering-countries/en-us/quality-education/preventing-school-dropouts-using-ml-and-analytics/.

2009. Certainly, subjects in survival analysis are usually followed over a specified period of time and the focus is on the time at which the event of interest occurs (Li et al., 2016). Thus, the benefit of using survival analysis over other methods is the ability to add the time component into the model and also effectively handle censored data. In spite of the success of survival analysis methods in other domains such as health care, engineering, etc., there is only a limited attempt of using these methods in student retention problem (Bani and Haji, 2017).

Matrix factorization is a clustering machine learning methods that can accommodate framework with some variations (Yang et al., 2014). The study presented by Hu and Rangwala (2017); Elbadrawy et al. (2016), described matrix factorization. In Elbadrawy et al. (2016) study, two classes of methods for building the prediction models were presented. The aim of the conducted study was to facilitate a degree planning and determine who might be at risk of failing or dropping a class. The first class builds models using linear regression approaches and the second class used matrix factorization approaches. Regression-based methods describe course-specific regression (CSpR) and personalized linear multi-regression (PLMR) while matrix factorization based methods associate standard Matrix Factorization (MF) approach. The mentioned approach was applied on the dataset generated from George Mason University (GMU) transcript data, University of Minnesota (UMN) transcript data, UMN LMS data, and Stanford University MOOC data. One limitation of the standard MF method is that, it ignores the sequence in which the students have taken the various courses. Besides, the latent representation of a course can potentially be influenced by the performance of the students in courses that were taken afterward.

Furthermore, the work present in Iam-On and Boongoen (2017) study, proposed a new data transformation model, which is built upon the summarized data matrix of link-based cluster ensembles (LCE). The aim of the conducted study was to establish the clustering approach as a practical guideline for exploring student categories and characteristics. This was accomplished using educational dataset obtained from the operational database system at Mae Fah Luang University, Chiang Rai, Thailand. Like several existing dimension reduction techniques such as Principal Component Analysis (PCA) and Kernel Principal Component Analysis (KPCA), this method aims to achieve high classification accuracy by transforming the original data to a new form. However, the common limitation of these new techniques is the demanding time complexity, such that it may not scale up well to a very large dataset. Whilst worst-Case Traversal Time (WCT-T) is not quite for a highly time-critical application, it can be an attractive candidate for those quality-led works, such as the identification of those students at risk of under achievement.

Deep neural network (DNN) is an approach based on Artificial Neural Networks (ANN) with multiple hidden layers between the input and output layers (Deng and Yu, 2014). While, Probabilistic Graphical Model (PGM) combine probability theory and graph theory to offer a compact graph-based representation of joint probability distributions exploiting conditional independences among the random variables (Pernkopf et al., 2013). Similar to shallow ANNs, DNNs can model complex non-linear relationships (Mun et al., 2017; Ramachandra and Way, 2018). Different deep learning architecture such as Recurrent Neural Network (RNN) and other probabilistic graphical model such as Hidden Markov Model (HMM) have been employed on the problem of student dropout (Fei and Yeung 2015).

The study presented by Fei and Yeung (2015), considered two temporal models which are state space models and recurrent neural networks. These approaches were applied in two MOOCs datasets, one offered on the Coursera platform, called "The Science of Gastronomy", and the other on the edX platform, called "Introduction to Java Programming". The aim of the conducted study was to identify students at risk of dropping out. State space models describe two variants of Input Output Hidden Markov Model (IOHMM) with continuous state space while, recurrent neural networks describe vanilla RNN and RNN with Long Short Term Memory (LSTM) cells as hidden units. IOHMM was proposed by for learning problems involving sequentially structured data. As it was originated from HMM, it learned to map input sequences to output sequences. Moreover, unlike the standard discrete-state HMM, the state space in described IOHMM formulation is continuous so the state space can bear more representation power compared with enumerating discrete states. Furthermore, Vanilla Recurrent Neural Network (Vanilla RNN), unlike feed forward neural networks such as the Multi Layer Perceptron (MLP), allows the network connections to form cycles.

The limitation of the conducted study was vanishing gradient problem. While an important property of RNNs is their ability to use contextual information in learning the mapping between the input and output sequences. A subtlety is that, for basic RNN models, the range of temporality that can be accessed in practice is usually quite limited so that the dynamic states of RNNs are considered as short term memory. This is because of the influence of a given input on the hidden layer. Therefore, on the network output it will either decays or blows up exponentially so as to cycles around the recurrent connections. To handle short-term

memory of RNNs last for longer so as to tackle the vanishing gradient problem, Long Short-Term Memory RNN (LSTM Network) was introduced.

On addressing the problem of student dropout, machine learning techniques have been applied in various platforms such as Massive Open On-line Course (MOOC) (Chen et al., 2017; Liang et al., 2016; Fei and Yeung 2015; Prieto et el., 2017) and other Learning Management System (LMS) such as Moodle (Elbadrawy et al., 2016; Hung et al., 2017; Santana et al., 2015). These platforms generated datasets which contain information that can be categorized into academic performance, socio-economic and personal information (Lei and Li, 2015). MOOC platforms such as Coursera and edX is among popular used platforms for generating datasets to be used in student dropout prediction (Chen et al., 2017). While, Moodle as a popular Learning Management System (Santana et al., 2015), provides public datasets such as UMN LMS (Elbadrawy et al., 2016). Furthermore, on identifying at risk students for early interventions, other researchers collected data from an on-line graduate program in the United States and validation was conducted by using Fall 2014 data set (Hung et al., 2017).

5 Open Challenges for Future Research

On previous sections we have presented an overview of machine learning techniques on addressing student dropout problem and highlighting the gaps and limitations. Despite several efforts done by previous researchers, there are still some challenges which need to be addressed.

It has been observed that, most of the algorithms have been developed and tested in developed countries using existing datasets generated from developed countries. Furthermore, MOOC and Moodle are among the most used platforms which offer public datasets to be used on addressing the student dropout problem. The limitation of public datasets from developing countries (Mgala and Mbogho, 2015), brought the need to develop more datasets from different geographical location. This may include transforming registration information of students with ongoing academic progress from paper based approach into electronic storage. However, cost and time must be acquired to accommodate the process. Furthermore, to the knowledge of researchers, there are only few researches which has been conducted in developing countries. Thus, further research is needed to explore the value of machine learning algorithms in cubing dropout in the context of developing countries with inclusion of factors that applied in the scenario.

Second, most of the presented works have focused on providing early prediction only (Lakkaraju et al., 2015). Therefore, developing countries' research should focus on facilitating a more robust and comprehensive early warning systems for students dropout which can identify students at risk in future cohorts (early warning mechanism), rank students according to their probability of dropping (ranking mechanism) and identifying students who are at risk even before they drop (forecasting mechanism).

Third, most existing studies ignore the fact that dropout rate is often low in existing datasets. This is a serious problem especially in the context of student retention (Thammasiri et al., 2014), with dropout students significantly less than those who stay and thus future research should consider developing a student dropout algorithm with consideration of data imbalance problem.

Fourth, many studies focus on addressing student dropout using student level datasets. However, developing countries need to include school level datasets on addressing this problem due to the issue of limited resources which face many school districts (Lakkaraju et al., 2015). This will involve the use of new sources school level data, that will consider school needs related features and applying additional machine learning approaches to improve predictive power of the proposed algorithm. The algorithm will enable relevant authorities to plan effectively and accurately, formulate policies, and make decisions on measures to address the problem; with concern of school level factors such as Pupil Teacher Ratio (PTR) which can be monitored by the authorities.

6 Conclusions

A survey of machine learning techniques on addressing student dropout problem is presented. The survey draws several conclusions; First, while several techniques have been proposed for addressing student dropout in developed countries, there is lack of research on the use of machine learning for addressing this problem in developing countries. Second, despite the major efforts on using machine learning in education, data imbalance problem has been ignored by many researchers. This facilitate using improper evaluation metrics on analyzing performance of the algorithms. Third, many researches focus on providing early prediction rather than including ranking and forecasting mechanisms on addressing the problem of student dropout. Lastly, school level datasets must be considered when addressing this problem, in order to come up with the proposed solutions to facilitate the authorities on identifying at risk schools for early intervention.

Acknowledgements
The authors would like to thank the African Development Bank (AfDB), Data for Local Impact (DLi) and Eagle Analytics company for supporting this study.

Competing Interests
The authors have no competing interests to declare.

References
Adhatrao, K, Gaykar, A, Dhawan, A, Jha, R and **Honrao, V.** 2013. Predicting Students' Performance Using Id3 and C4.5 Classification Algorithms. *International Journal of Data Mining and Knowledge Management Process*, 3(5): 39–52. DOI: https://doi.org/10.5121/ijdkp.2013.3504

Ameri, S. 2015. *Survival Analysis Approach For Early Prediction Of Student Dropout*. PhD thesis, Wayne State University.

Ameri, S, Fard, MJ, Chinnam, RB and **Reddy, CK.** 2016. Survival Analysis Based Framework for Early Prediction of Student Dropouts. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM)*, 16: 903–912. New York, NY, USA: ACM. DOI: https://doi.org/10.1145/2983323.2983351

Arsad, PM, Buniyamin, N and **Manan, J-lA.** 2013. A neural network students' performance prediction model (NNSPPM). *2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, (November): 1–5.

Aulck, L, Aras, R, Li, L, Heureux, CL, Lu, P and **West, J.** 2017. STEM-ming the Tide: Predicting STEM attrition using student transcript data. In: *Proceedings of ACM Knowledge Discovery and Data Mining Conference*. Nova Scotia, Canada.

Aulck, L, Velagapudi, N, Blumenstock, J and **West, J.** 2016. Predicting Student Dropout in Higher Education. In: *ICML Workshop on #Data4Good: Machine Learning in within the Open Polytechnic of New Zealand, relying Social Good Applications*. New York, NY, USA.

Babu, AR. 2015. Comparative Analysis of Cascadeded Multilevel Inverter for Phase Disposition and Phase Shift Carrier PWM for Different Load. *Indian Journal of Science and Technology*, 8(April): 251–262. DOI: https://doi.org/10.17485/ijst/2015/v8iS7/70151

Bani, MJ and **Haji, M.** 2017. College Student Retention: When Do We Losing Them? In: *Proceedings of the World Congress on Engineering and Computer Science*. Tehran, IRAN.

Beck, HP and **Davidson, WD.** 2016. Establishing an Early Warning System: Predicting Low Grades in College Students from Survey of Academic Orientations ... *Research in Higher Education*, 42(December 2001).

Brundage, A. 2014. The use of early warning systems to promote success for all students.

Center for Digital Technology and Management. 2015. The Future of Education Trend Report 2015. *Technical report*, Munich, Germany.

Chen, JF, Hsieh, HN and **Do, QH.** 2014. Predicting student academic performance: A comparison of two meta-heuristic algorithms inspired by cuckoo birds for training neural networks. *Algorithms*, 7(4): 538–553. DOI: https://doi.org/10.3390/a7040538

Chen, Y, Chen, Q, Zhao, M, Boyer, S, Veeramachaneni, K and **Qu, H.** 2017. DropoutSeer: Visualizing learning patterns in Massive Open Online Courses for dropout reasoning and prediction. *2016 IEEE Conference on Visual Analytics Science and Technology, VAST 2016 – Proceedings*, 111–120.

Deng, L and **Yu, D.** 2014. Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, 7(3–4): 197–387.

Durairaj, M and **Vijitha, C.** 2014. Educational data mining for prediction of student performance using clustering algorithms. *International Journal of Computer Science and Information Technologies (IJCSIT)*, 5(4): 5987–5991.

Elbadrawy, A, Polyzou, A, Ren, Z, Sweeney, M, Karypis, G and **Rangwala, H.** 2016. -okay-Predicting Student Performance Using Personalized Analytics. *Computer*, 49(4): 61–69. DOI: https://doi.org/10.1109/MC.2016.119

Erik, G. 2014. Introduction to Supervised Learning.

Fei, M and **Yeung, D-Y.** 2015. Temporal Models for Predicting Student Dropout in Massive Open Online Courses. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 256–263. DOI: https://doi.org/10.1109/ICDMW.2015.174

Gao, T. 2015. *Hybrid classification approach of SMOTE and instance selection for imbalanced datasets*. PhD thesis, Iowa State University.

Gray, G, McGuinness, C and **Owende, P.** 2014. An application of classification models to predict learner progression in tertiary education. *2014 4th IEEE International Advance Computing Conference (IACC)*, 549–554. DOI: https://doi.org/10.1109/IAdCC.2014.6779384

Halland, R, Igel, C and **Alstrup, S.** 2015. High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study. *ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, (April): 22–24.

Hu, Q and **Rangwala, H.** 2017. Enriching Course-Specific Regression Models with Content Features for Grade Prediction. In: *Proceedings of ACM SIGKDD*. Nova Scotia, Canada. DOI: https://doi.org/10.1109/DSAA.2017.74

Hung, JL, Wang, MC, Wang, S, Abdelrasoul, M, Li, Y and **He, W.** 2017. Identifying At-Risk Students for Early Interventions – A Time-Series Clustering Approach. *IEEE Transactions on Emerging Topics in Computing*, 5(1): 45–55. DOI: https://doi.org/10.1109/TETC.2015.2504239

Iam-On, N and **Boongoen, T.** 2017. Generating descriptive model for student dropout: A review of clustering approach. *Human-centric Computing and Information Sciences*, 7(1): 1. DOI: https://doi.org/10.1186/s13673-016-0083-0

Iqbal, Z, Qadir, J, Mian, AN and **Kamiran, F.** 2017. Machine Learning Based Student Grade Prediction: A Case Study.

Jordan, MI and **Mitchell, TM.** 2015. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245): 255–260. DOI: https://doi.org/10.1126/science.aaa8415

Joseph, HR. 2014. Promoting education: A state of the art machine learning framework for feedback and monitoring E-Learning impact. *2014 IEEE Global Humanitarian Technology Conference – South Asia Satellite, GHTC-SAS 2014*, 251–254. DOI: https://doi.org/10.1109/GHTC-SAS.2014.6967592

Kartal, OO. 2015. *Using Survival Analysis to Investigate the Persistence of Students in an Introductory Information Technology Course at Metu*. PhD thesis, The Middle East Technical University.

Kotsiantis, SB. 2012. Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37(4): 331–344. DOI: https://doi.org/10.1007/s10462-011-9234-x

Kumar, M, Singh, AJ and **Handa, D.** 2017. Literature Survey on Educational Dropout Prediction. *I.J. Education and Management Engineering*, 2(March): 8–19. DOI: https://doi.org/10.5815/ijeme.2017.02.02

Lakkaraju, H, Aguiar, E, Shan, C, Miller, D, Bhanpuri, N, Ghani, R and **Addison, KL.** 2015. A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. *KDD*, 1909–1918. DOI: https://doi.org/10.1145/2783258.2788620

Lan, AS, Studer, C and **Baraniuk, RG.** 2014. Time-varying Learning and Content Analytics via Sparse Factor Analysis. In: *KDD'14 ACM*. New York, USA. DOI: https://doi.org/10.1145/2623330.2623631

Latif, A, Choudhary, AI and **Hammayun, AA.** 2015. Economic Effects of Student Dropouts: A Comparative Study. *Journal of Global Economics*, 03(02): 2–5.

Lee, K. 2017. Large-Scale and Interpretable Collaborative Filtering for Educational Data.

Lei, C and **Li, KF.** 2015. Academic Performance Predictors. In: *Proceedings – IEEE 29th International Conference on Advanced Information Networking and Applications Workshops, WAINA 2015*. DOI: https://doi.org/10.1109/WAINA.2015.114

Li, Y, Wang, J, Ye, J and **Reddy, CK.** 2016. A Multi-Task Learning Formulation for Survival Analysis. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD'16*, 1715–1724. DOI: https://doi.org/10.1145/2939672.2939857

Liang, J, Li, C and **Zheng, L.** 2016. Machine learning application in MOOCs: Dropout prediction. *ICCSE 2016 – 11th International Conference on Computer Science and Education (ICCSE)*, 52–57.

Lin, WJ and **Chen, JJ.** 2013. Class-imbalanced classifiers for high-dimensional data. *Briefings in Bioinformatics*, 14(1): 13–26. DOI: https://doi.org/10.1093/bib/bbs006

Longadge, R, Dongre, SS and **Malik, L.** 2013. Class imbalance problem in data mining: Review. *International Journal of Computer Science and Network*, 2(1): 83–87.

López, V, Fernández, A, García, S, Palade, V and **Herrera, F.** 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250: 113–141. DOI: https://doi.org/10.1016/j.ins.2013.07.007

Mgala, M. 2016. *Investigating Prediction Modelling of Academic Performance for Students in Rural Schools in Kenya*. PhD thesis, University of Cape Town.

Mgala, M and **Mbogho, A.** 2015. Data-driven Intervention-level Prediction Modeling for Academic Performance. *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development*, 2: 1–8. DOI: https://doi.org/10.1145/2737856.2738012

Młynarska, E, Greene, D and **Cunningham, P.** 2016. Time series clustering of Moodle activity data. *CEUR Workshop Proceedings*, 1751: 104–115.

Mosha, D. 2014. *Assessment of Factors behind Dropout in Secondary Schools in Tanzania. A Case of Meru District in Tanzania*. PhD thesis, Open University of Tanzania.

Mun, S, Shin, M, Shon, S, Kim, W, Han, D and **Ko, H.** 2017. DNN transfer learning based non-linear feature extraction for acoustic event classification. *IEICE Transactions on Information and Systems*, E100D(9): 1–4. DOI: https://doi.org/10.1587/transinf.2017EDL8048

Natek, S and **Zwilling, M.** 2014. Expert Systems with Applications Student data mining solution knowledge management system related to higher education institutions. *Expert Systems with Applications*, 41: 6400–6407. DOI: https://doi.org/10.1016/j.eswa.2014.04.024

Nunn, S, Avella, JT, Kanai, T and **Kebritchi, M.** 2016. Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review. *Online Learning*, 20(2): 13–29. DOI: https://doi.org/10.24059/olj.v20i2.790

Patron, R. 2014. Early school dropouts in developing countries: An equity issue? The Uruguayan case. *University of Uruguay*, P13.

Pernkopf, F, Peharz, R and **Tschiatschek, S.** 2013. *Introduction to Probabilistic Graphical Models Introduction*. Graz, Austria.

Poh, N and **Smythe, I.** 2015. To what extend can we predict students' performance? A case study in colleges in South Africa. *IEEE SSCI 2014–2014 IEEE Symposium Series on Computational Intelligence – CIDM 2014: 2014 IEEE Symposium on Computational Intelligence and Data Mining, Proceedings*, 416–421.

President's Office and Government, Regional Administration and Local. 2016. Pre-Primary, Primary and Secondary Education Statistics in Brief 2016 The United Republic of Tanzania President's Office Regional Administration and Local Government. *Technical report*.

Prieto, LP, Rodríguez-Triana, MJ, Kusmin, M and **Laanpere, M.** 2017. Smart school multimodal dataset and challenges. *CEUR Workshop Proceedings*, 1828: 53–59.

Rakesh, A, Christoforaki, M, Gollapudi, S, Kannan, A, Kenthapad, K and **Swaminathan, A.** 2014. Mining Videos from the Web for Electronic Textbooks. *Microsoft Research*.

Ramachandra, V and **Way, K.** 2018. Deep Learning for Causal Inference.

Rovira, S, Puertas, E and **Igual, L.** 2017. Data-driven system to predict academic grades and dropout. *PLOS ONE*, 12(2): 1–21. DOI: https://doi.org/10.1371/journal.pone.0171207

Sales, A, Balby, L and **Cajueiro, A.** 2016. Exploiting Academic Records for Predicting Student Drop Out: a case study in Brazilian higher education. *Journal of Information and Data Management*, 7(2): 166–180.

Santana, MA, Costa, EB, Neto, BFS, Silva, ICL and **Rego, JBA.** 2015. A predictive model for identifying students with dropout profiles in online courses. *CEUR Workshop Proceedings*, 1446.

Sathya, R and **Abraham, A.** 2013. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2): 34–38. DOI: https://doi.org/10.14569/IJARAI.2013.020206

Shahidul, SM and **Karim, AHMZ.** 2015. Factors contributing to school dropout among the girls: a review of literature. *European Journal of Research and Reflection in Educational Sciences*, 3(2): 25–36.

Shahiri, AM, Husain, W and **Rashid, NA.** 2015. A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, 72: 414–422. DOI: https://doi.org/10.1016/j.procs.2015.12.157

TAMISEMI. 2004. The United Republic of Tanzania Ministry of Education and Culture. 2004–2009.

Thammasiri, D, Delen, D, Meesad, P and **Kasap, N.** 2014. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, 41(2): 321–330. DOI: https://doi.org/10.1016/j.eswa.2013.07.046

UNESCO. 2011. UNESCO Global Partnership for Girls' and Women's Education- One Year On.

US Department of Education. 2016. Definition of Early Warning Systems Research on Early Warning Systems Issue Brief: Early Warning Systems. *Technical Report* September.

Wang, P, Li, Y and **Reddy, CK.** 2017a. Machine Learning for Survival Analysis: A Survey. *ACM Comput. Surv. Article*, 1(1): 38.

Wang, W, Yu, H and **Miao, C.** 2017b. Deep Model for Dropout Prediction in MOOCs. *Proceedings of the 2nd International Conference on Crowd Science and Engineering – ICCSE'17*, 26–32. DOI: https://doi.org/10.1145/3126973.3126990

Waters, AE, Studer, C and **Baraniuk, RG.** 2014. Sparse Factor Analysis for Learning and Content Analytics. *Journal of Machine Learning Research*, 15: 1959–2008.

Xu, J, Moon, KH and **van der Schaar, M.** 2017. A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs. *IEEE Journal of Selected Topics in Signal Processing*, 11(5): 742–753. DOI: https://doi.org/10.1109/JSTSP.2017.2692560

Yang, D, Piergallini, M, Howley, I and **Rose, C.** 2014. Forum Thread Recommendation for Massive Open Online Courses. *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*, 257–260.

Yudelson, MV, Koedinger, KR and **Gordon, GJ.** 2013. Individualized Bayesian Knowledge Tracing Models.

How to cite this article: Mduma, N, Kalegele, K and Machuve, D. 2019. A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction. *Data Science Journal,* 8: 14, pp. 1–10. DOI: https://doi.org/10.5334/dsj-2019-014

Submitted: 04 October 2018 **Accepted:** 19 March 2019 **Published:** 17 April 2019

Copyright: © 2019 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS