School dropout profiling and prediction approach using machine learning

Khamisi Kalegele

Department of ICT,
Open University of Tanzania (OUT),
Dar es Salaam, Tanzania
Email: kalegs03@gmail.com
Email: khamisi.kalegele@out.ac.tz

Abstract: Tens of thousands of children drop out of schools in Sub-Saharan Africa while the widely adopted interventions are either reactive in nature or uninformed. The increasing availability of disaggregated data has opened new prospects for proactive interventions using new data technologies in unprecedented ways. However, awareness and skills among wider interest groups including school managers, researchers and developers are at staggering low levels. This paper demonstrates how a machine learning approach can be used to profile students and predict the likelihood of dropping out of school in order to enable proactive interventions and potentially inform youth related policies. Using well known open dataset, supervised and unsupervised profiling approaches are demonstrated, compared and proved to be better performers than a traditional approach. The approaches can be replicated under real production environment using actual data to enable informed interventions and reduce dropouts.

Keywords: school dropout; profiling; prediction; truancy; youth empowerment; machine learning; classification; data-driven; absenteeism; Africa.

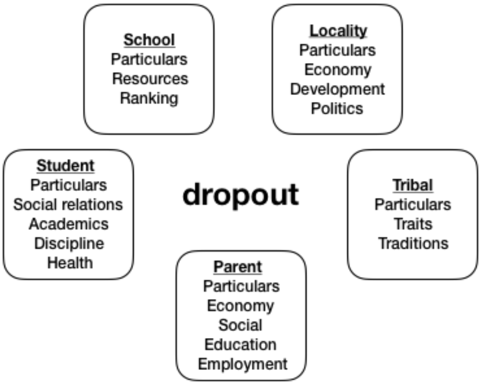
Reference to this paper should be made as follows: Kalegele, K. (2020) ‘School dropout profiling and prediction approach using machine learning’, *Int. J. Information Technology, Communications and Convergence*, Vol. 3, No. 4, pp.245–258.

Biographical notes: Khamisi Kalegele holds a PhD in Computer/Information Sciences from Tohoku University in Sendai, Japan. Currently, he is a Senior Lecturer at the Open University of Tanzania. Previously, he was the Chief Research Officer at the Tanzania Commission for Science and Technology between 2016–2019; a Lecturer at the Nelson Mandela African Institution of Science and Technology; and a Senior Radio Frequency Engineer at Celtel Tanzania Limited between 2003–2007. His research interest is on data-driven and machine learning approaches for socio-economic sectors of developing countries.

1 Introduction

While the youthful population, like natural resources, forms the backbone of Africa's attractiveness, the population's unproductiveness has become a huge threat to future development. Various reports outline some of the key underlying factors to include lack of well formulated youth policies for promotion of skills development, and science and technology (UNEconomicCommissionAfrica, 2017; ACBF, 2017) In particular, a significant percentage of youth are under a threat of not completing school due to various reasons including death, pregnancy, and truancy. When young students drop out of schools, there are elevated risks of increase in criminal activities, unemployment, drugs and other similar adversities (Thornberry et al., 1985). Such risks are associated with huge economic costs and are detrimental to social capital (Theunissen et al., 2015). In a case of Tanzania, about 130,000 students dropped out of primary and secondary schools annually between 2010 and 2016 (PORALG, 2016). There has been inadequate disaggregation of data to enable identification of root causes, as a result, most of the dropouts are generally attributed to death, pregnancy and truancy only. Literature has confirmed a number of determinants that relate to dropout with notable variations across regions. Despite of the known determinants, interventions have always been reactive in nature and uninformed, for instance, family outreach and affective interventions targeting even those who are not in danger. The aim of this study is to offer insights into practical modelling of determinants of dropout using machine learning techniques in order to enable implementation of early warning systems. This will in turn extend the local use of data in proactive management of dropouts, particularly in designing interventions. Globally, dropout interventions have proven to be quite complicated. In the Netherlands, for instance, three randomised controlled trials using a bottom-up approach to reduce dropouts in vocational colleges did not significantly succeed (Bolhaar et al., 2019). In their study, it was analysed that the interventions (renewed intake, absenteeism counsellors and e-coaching) influenced unauthorised absenteeism only, which was not an absolute determinant.

Figure 1 Example attributes for typical dropout determinants



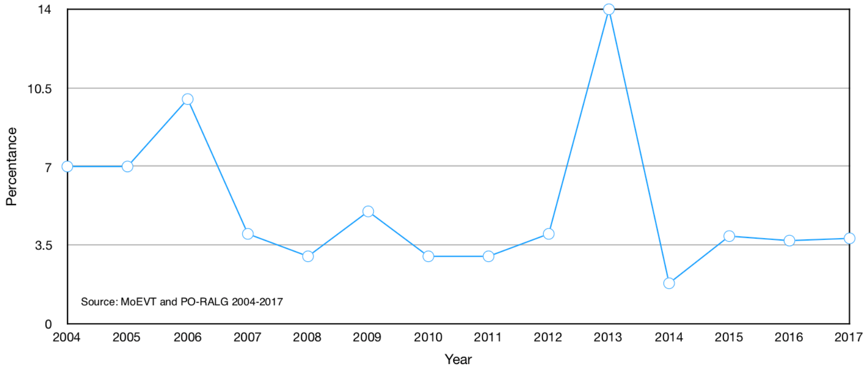
The determinants that influence dropouts, as presented in Theunissen et al. (2015), Lee and Chung (2019), Tan and Shao (2015) and Sivakumar et al. (2016), can be categorised based on their relationship to school environment, student's personal attributes, student's health, academics, parents, ethnicity or tribal, and locality. In machine learning representation, an instance t of these determinants relates to an outcome $y = \{dropout, not\ dropout\}$. Given a set of t, the traditional challenge has been to building a model for classification or prediction as accurately as possible. Existing dropout prediction studies have focused on differing aspects including prediction performance by attempting to balance datasets, improving algorithms, and selection of determinant's attributes (Sivakumar et al., 2016; Lee and Chung, 2019; Tan and Shao, 2015). In many settings, especially in low to middle income countries, the chief stumbling block for the practical application of machine learning to predict dropouts and monitor relevant determinants has been the lack and ineffectiveness of data systems. The key attributes of dropout determinants, shown in Figure 1, are not sufficiently collected to warrant practical use of advanced data technologies. Over the recent years, significant strides to computerise systems in critical sectors such as health and education have been made. In Tanzania, a multitude of electronic systems are being put into use to improve effectiveness and performance of educational data systems. However, there are still issues on data quality, levels of disaggregation from national level to schools, accessibility and use of data. This current situation weakens the premises for reducing dropouts and planning interventions through insights emanating from the relevant determinants. Nonetheless, there is enough indication that the available data can be used to capture common patterns of dropping students, and as more data become available, updates can be done. Practicability of dropout prediction models derived from impartial set of characterising attributes can be increased by the fact that the nature of the problem requires capturing of all potential dropouts. It can be argued that non-dropping students would rather be suspected to drop that the opposite. Therefore, it is desirable to develop models with a careful designed performance metric.

Machine learning has come of age as a tool for building practical prediction models for transforming early warning systems in schools. The associated techniques and methods such as batch learning, incremental learning and online learning offer a wide range of choices to fit specific contexts and scenarios. In this article, a specific approach to apply machine learning techniques to monitor dropoutness is proposed. Dropoutness is the likelihood of a student to drop out of school as per local rules and regulation. As public awareness to various technologies increases steadily, this paper serves to provide inspiration for wider deployment of effective early warning system in managing dropouts.

1.1 Dropout in Tanzania: case study

The 2012 national census revealed that 50% of Tanzanian population is below the age of 18 years making Tanzania one of the ten countries with biggest population of youth in the world (Restless Development, 2011). The government has since made deliberate efforts to improve access to education as part of its effort to leverage the youth base for future economic growth and prosperity. Despite the efforts, statistics indicate that 85985 and 61488 pupils dropped out of primary and secondary schools respectively in 2015 (PORALG, 2016). Figure 2 shows percentages of dropouts in Tanzania from 2004 to 2017 depicting that dropouts are at unacceptable levels. Dropout occurrence is largely due to adolescence challenges, and partly due to death and social-economic challenges. The fight against dropouts is of paramount importance to both the central government and local governments. Authorities at various governance levels have for sometime now understood how imperative is the collection of relevant education management data from school level. Therefore, a lot of success has been made whereby vital education statistics are now electronically integrated from district levels to the national level (PORALG, 2016). The ongoing government efforts indicate that there will soon be electronic systems for collection of statistics from school level. This article is part of the efforts to strengthen sector preparedness from the perspective of data use, research, and informed problem interventions. The article presents an ICT for development (ICT4D) research on dropout prediction in order to ensure broader contribution to relevant development challenges.

Figure 2 Trend of dropout rates in Tanzania (see online version for colours)



Source: MoEVT and PO-RALG 2004-2017

2 Related works

Dropout as a phenomenon is common in many settings such as schools, colleges, and e-learning programmes (Jia et al., 2015; Dupere et al., 2017; Tan and Shao, 2015). While researchers in the developed world (Lee and Chung, 2019; Tan and Shao, 2015) have already attempted to design machine learning approaches to combat the challenge, efforts in developing countries (PORALG, 2016) are largely manual and compounded by reactiveness. This study contributes to efforts of eliminating dropouts in secondary and primary schools in developing countries by introducing proactiveness and enabling early warning. A key contextual difference is the extent of digitalisation of information management in schools whereas the situation in developing countries is far below. The implication here is that there is no sufficient data available to undertake quality studies. Existing dropout prediction studies in developed countries have focused on improving learning performance using recorded data, balancing available datasets and comparing algorithms (Sivakumar et al., 2016; Lee and Chung, 2019; Tan and Shao, 2015). Another pocket of researches from the global body of knowledge reported approaches to predict dropout from courses which run online over academic semesters (Márquez et al., 2016; Whitehill et al., 2017). Significantly, these differ from dropout situation in secondary schools in developing countries primarily due to shortage of data making the approaches impractical. Approaches that predict dropout in online education programmes (including massive open online course – MOOC) deals technically with adult or matured professionals and therefore have a different set of complexities (Choi and Park, 2018; Liang et al., 2016). For instance, in Tanzania, secondary and primary education is mandatory and enforced by law whereby the responsibility lies with guardians. The approach presented in this paper aims at enabling practical implementation of early warning systems in the context of developing countries where data is neither always available for all students nor in acceptable quality. The approach involves the use of simple but practical methods to profile students and subsequently predict the likelihood of dropping out of school.

In a recent review work by Lwoga and Sangeda (2019), various reviews were systematically synthesised and it was found out that evidence on the contribution of ICT in addressing developmental challenges is limited. This is partly attributed to unhealthy diversity of project objectives which in turn affects their design, implementation and outcomes. The work presented here serves to inspire further research specifically on dropout prediction models so as to provide enough validated options to inform policy processes and aid the fight to eliminate dropouts.

3 Methods

3.1 Data

Three datasets were considered for the study:

1 Government data published by the President's Office Regional Administration and Local Government (PORALG).

2 Uwezo survey data from Twaweza Civil Society Organization (Twaweza, 2015).

3 Student performance dataset (Cortez and Silva, 2008).

The government data is an aggregation of daily attendance data from every school of Tanzania through the Basic Education Management Information System (BEMIS). The data comprises of school details and number of dropouts by reasons, i.e., death, pregnancies, and truancy. The Uwezo survey is an annual survey that gathered information about students from schools, households and villages. It included the following fields - area, household details like size and number of meals, main source of household income, students particulars like age, availability of girls room in schools, number of teachers, and parent-teacher meetings. The student performance data relates to achievement in secondary education of two schools (Cortez and Silva, 2008). Through school reports and questionnaires, key attributes (student grades, demographic, social and school related features) were collected. After analysing the datasets, student performance dataset presented desired level of attribute richness to be preferred for the study. As summarised in Table 1, the selected dataset has more attributes than the other two datasets.

Table 1 Availablity of key attributes in datasets

Attribute	*Government published*	*Twaweza*	*Student Opendataset*
School			
Particulars	✓	✓	
Resources	✓	✓	
Performance	✓		
Student			
Particulars		✓	✓
Social relations			✓
Academics			✓
Discipline			
Health			✓
Parents			
Particulars			✓
Economy		✓	✓
Social			
Education			✓
Job status			✓
Locality			
Particulars			✓
Economy			
Development			
Political			

3.2 Processing

The student performance dataset contains data items that are not labelled but has the total number of days a student was absent. Using the number of absent days, the data items were labelled as *dropout* or *not dropout* based on local rules and regulations. Dropout can be defined as the failure to attend school for a specified number of days (in Tanzania, it is 90 days). Moreover, irrelevant attributes such as address were removed from the dataset. The remaining attributes were sex, age, family size, parent status, mother education, father education, mother job, father job, reason for joining school, travel time, study time, number of failures, family support, extra paid classes, extra curricular activities, nursery school attendance, aspiration for joining higher education, internet, romantic relationship, family relations, free time, going out and health condition. All processing was done using R statistical computing tool.

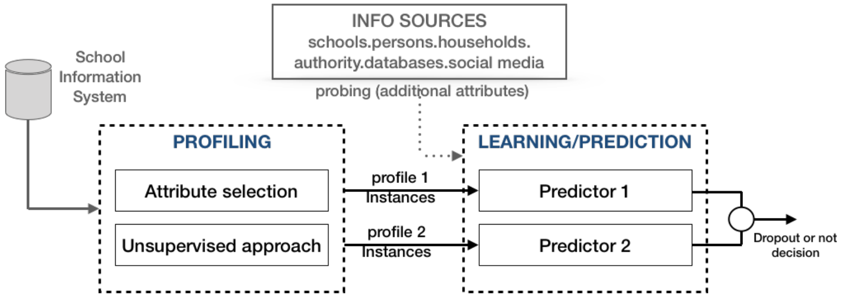
3.3 Profiling, learning and prediction

The used approach involves three stages of students profiling, digital probing, and dropout prediction as shown in Figure 3. During profiling stage, two approaches (*attribute selection* and *unsupervised learning*) are proposed for putting students into similar groups (herein referred to as profiles). Profiling students is typical of many approaches dealing with massive students data such as online tutoring systems and blended learning (Harrak et al., 2018; Bouchet et al., 2013). The process mitigates the undesirable computational overhead risks imposed onto computing infrastructure of schools and organisations. In the second stage, additional information about riskier students is either probed or harvested from available sources such as social media pages and national databases. This additional information helps to move individual students across profiles and to improve prediction performance in the next stage. Lastly, the actual prediction about whether the student is dropping or not is done preceded by model learning from training data. Specific methods for each of these stages are presented in the next paragraphs.

$$IG(T, a) = H(T) - H(T/a) \tag{1}$$

In schools, there are students whose likelihood of dropping is minimally small. The stage of profiling students aims at identifying these students so that more resources are directed to those who are more likely to drop, i.e., minimising false negatives. Statistically, it is possible to identify determinants for students who are unlikely to drop provided there is enough observations to support it. A simple method of relying on information gain is used because there is no enough observations to enable more sophisticated approaches (Quinlan, 1986). Given by equation (1), Information gain measures how much information, about dropout possibility, a specific attribute gives. In the equation, $H(T/a)$ is the conditional entropy of T given the value of attribute a. The top h attributes with higher information gain are then selected to develop a simple decision tree for profiling students. For instance, if it is generally observed that girls from a certain province s tend to drop out of school in relatively large numbers, then a simple decision tree is used to classify students into *profile 1* (girls of province s) and *profile 2* (the rest of students). Note that the attributes *gender* = '*female*' and *area* = '*province* s' (used to profile students) are only a subset of attributes required for predicting dropout. Through a trial and error process, led by a domain expert, the value of h is determined to be the one that offers the best tree learning performance without constraining computational resources at a school.

Figure 3 A block diagram showing the approach using two profiles (see online version for colours)



Another approach of using k subgroups of students that are deemed similar based on partial proximity (or dissimilarity) was also tested. This later approach is unsupervised while the former profiler approach is supervised. For this unsupervised profiler, Gower distance is selected to enable the computation of measures of proximity between profiles of students (Gower and Gower, 1971). This is a partitioning method that helps to define profiles of students in such a way that the overall intra-profile variation is minimum. In order to obtain an optimal number of profiles, average silhouette approach was used to determine how best each student fit within her/his group (Rousseeuw, 1987). A high average silhouette width indicates a good partitioning and therefore an optimal number of profiles. All computations were undertaken using R statistical and relevant packages (*dplyr* for data cleaning, *cluster* for clustering, and *ggplot2* for plotting).

Once the profiles are established, additional information attributes can be harvested in the second stage. In this study, the student performance dataset (Cortez and Silva, 2008) that is used, sufficiently contains a total of 33 attributes. There was no need for any particular probing or harvesting but numerous methods for probing have already been tested by many other researchers. Typical probing methods are empowered by:

1 social media APIs, such as Facebook Graph API and Instagram API, that can be used to harvest information

2 pop-ups which are normally used to gather or communicate information from application users (Huang and Kao, 2018).

Fekete (2020) compiled 15 examples of how popups are used to collect critical information from application users. A typical scenario would be – "everytime a parent visits a school website or student information system, a predefined question (e.g., that satisfies missing information in the database) pops up". Generally, these methods involve the use of mobile phones, websites, mobile apps, searching parent's Google profiles, and social media activities.

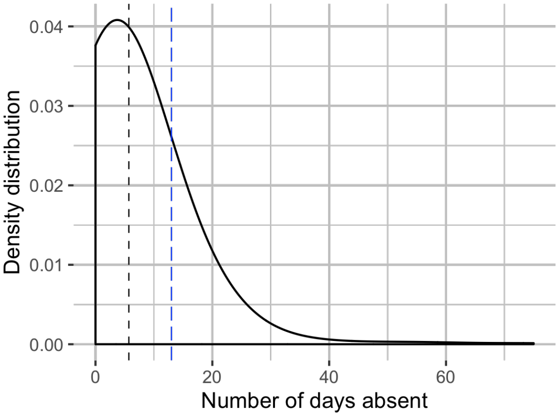
The final stage of dropout prediction is done using either 2 (if information gain is used) or k decision trees (if silhouette approach is used) in a custom ensemble fashion based on preceding student profiling. Since the type and performance of algorithm were no objective of this study, C4.5 algorithm was selected to build classification predictors for convenience (Hall et al., 2009). This study provides a performance comparison between *attribute selection* (i.e., supervised profiler), *unsupervised profiler* and *non-profiling* using commonly used point metrics of classification accuracy. Furthermore, due to the nature of the dropout problem and desired applicability of the solution, true positive rate (also known as recall) for the dropout class is also used for comparison. Based on intuition, it is more significant to ensure that dropout positive cases are actually classified as such in order to maximise true positive rates.

4 Results

4.1 Assigning labels

Dropout is determined by truancy for more than x number of days based on local rules, regulations and context. In the dataset, the maximum number of truancy is 93. Under the assumption that the reported truancy is continuous, distribution of dropout at varying dropout cut-off number was studied. For the sake of this study, dropout labels were assigned for student with truancy exceeding 13 just above a mean value as shown in Figure 4. The Figure shows density of number of absent days with a mean value indicated by the first dashed line (black colour) while the dropout cut-off number of 13 indicated by the second long-dashed line (blue colour). This resulted into 48 dropout cases out of a total of 395 students which is about 12%, sufficient enough for the purpose of the study.

Figure 4 Distribution of dropout after cut-off number of truancies



4.2 Profiling students

4.2.1 Attribute selection

From the results of information gain as explained in Section 3.3, the top four attributes were selected to build a predictor called the profiler with an average accuracy of 69.114%. The four attributes (age, reason for joining, average number of failures, and romantic relationship) were selected because they led to a meaningful predictor using simple-k-means approach as shown in Table 2. The table shows that students categorised in *profile 2* are more likely to drop than those in *profile 1* with dropout proportions of 21.970% and 7.224% respectively. That is, *profile 2* students are three times likely to drop than *profile 1* students.

Table 2 Selected attributes and resulting profiles

Attribute	*General Profile*	*Profile 1*	*Profile 2*
Age	16.696	16.418	17.250
Reason for joining school	attractive courses	attractive courses	near home
Average number of failures	0.334	0.190	0.621
Involvement in romantic relationship	no	no	yes
Number of students	395	263	132
Ratio of dropping students	12.151%	7.224%	21.970%

4.2.2 Silhouette width

The students performance data (Cortez and Silva, 2008) was clustered into two profiles using K-means algorithm. The decision to use two profiles is based on the fact that silhouette width for two profiles is the biggest as shown in Figure 5 (Amorim and Hennig, 2015; Rousseeuw, 1987). The resulting profiles of students were XY plotted as shown in Figure 6. The ratios of dropping students for the two profiles were found to be 12.621% and 11.640%.

Figure 5 Silhouette width for differing number of clusters

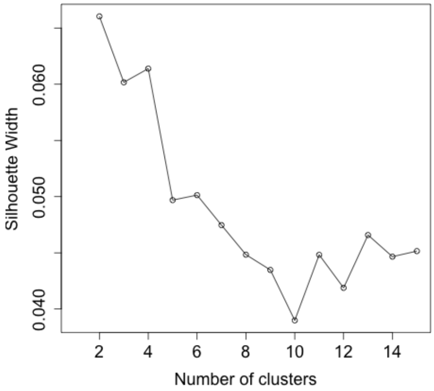
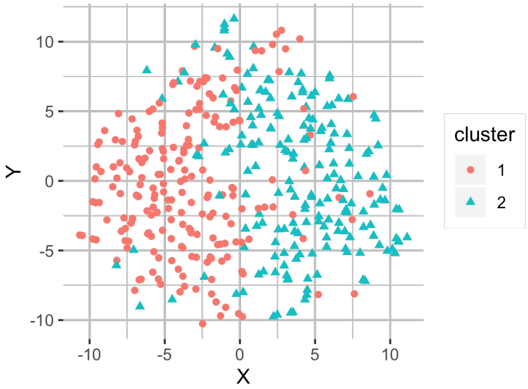


Figure 6 XY-plot of two clusters of students (see online version for colours)



4.3 *Learning and prediction*

The results for the two approaches (attribute selection and unsupervised silhouette profiling) that were investigated and the baseline traditional approach of using the whole dataset are presented in this subsection. Each of the two profiling approaches led to two decision tree predictors. Summary of performances of the approach in learning and prediction are shown in Tables 3 and 4, and discussed in the next section.

Table 3 Training performances

Metric	*Attribute selection*	*Unsupervised*	*Traditional*
Profile 1			
Accuracy	92.771	84.466	NA
True positive rate (dropout)	0.000	0.154	NA
Size of dataset	249	206	NA
Tree size (leaves)	1(1)	15(8)	NA
Profile 2			
Accuracy	75.343	87.302	NA
True positive rate (dropout)	0.200	0.000	NA
Size of dataset	146	189	NA
Tree size (leaves)	23(12)	25(13)	NA
Overall			
Accuracy (%)	NA	NA	85.570
True positive rate (dropout)	NA	NA	0.063
Size of dataset	NA	NA	395
Tree size (leaves)	NA	NA	1(1)

Table 4 Prediction performances

Metric	*Attribute selection*	*Unsupervised*	*Traditional*
Profile 1			
Accuracy	NA	88.354	NA
True positive rate (dropout)	NA	0.313	NA
Profile 2			
Accuracy	88.861	89.620	NA
True positive rate (dropout)	0.458	0.396	NA

5 Discussion

The results presented in Tables 3 and 4 depicts that profiling can actually improve both dropout learning and prediction performance. In this study, a limited test using a simple C4.5 algorithm and two metrics (accuracy and recall) is used to show how profiling can be done. Supervised profile 2, for instance, demonstrates highest prediction recall of 0.458 compared to unsupervised. Dropout learning is also more effective with supervised profile 2, as indicated by a recall of 0.200 higher than the second best, unsupervised profile 1 at 0.154 (about 25% performance difference). Further studies can be undertaken to assess various alternatives to profiling using different algorithms and datasets. Another important observation regarding learning performance is on the effect of parameter tuning and dataset preparation. From the tables, decision trees could not be learnt from the overall dataset and from supervised profile 1 as indicated by $size = 1$ and $leaves = 1$. This attest the importance of tuning learning parameters and improving learning datasets. For unsupervised profiling, as smaller the value of silhouette width is, the profiles still came out distinct enough as shown in Figure 5. This indicates viability of using unsupervised methods for dropout learning and prediction. Most studies end at pilot level, and are neither published nor disseminated to inform policy (Duncombe, 2016). To advance this work, efforts are underway to acquire good data of local context using a newly developed school attendance tool that will help to ascertain these results. It is believed that the performances of the predictors can be improved using ensemble methods and bootstrapping techniques (Kotsiantis, 2014).

Although the technique presented in this article is sufficiently systematic using information gain and machine learning algorithms, the resulting student's profiles can somehow appear to be arbitrary. As pointed out by Cherney and Price (2018), profiling-based efforts encourage focus on deviants (dropouts in this case) rather than paying attention to the underlying contextual and environmental factors. Therefore, enough attention will be needed to prevent stigmatisation of students with certain undesirable traits such as sex and tribe.

6 Conclusions

This article advocates the use of machine learning techniques to address the challenging problem of dropout in Tanzania and demonstrates how such techniques can be applied in a practical manner. Typically, application of machine learning is constrained by insufficiency of data. This article presented an approach that profiles students using available data and allows new data, as encountered, to be incorporated into learning dropout predictors. Using supervised and unsupervised alternatives, it has been shown that profiling students can be done to improve dropout learning and prediction processes, and also a provision for incorporation of new data into the processes. As discussed, additional studies are needed to further investigate profiling using other algorithms and also to further improve performance of both dropout profilers and predictors.

References

ACBF (2017) *Africa Capacity Report ACR 2017* [online] https://www.acbf-pact.org/what-we-do/how-we-do-it/knowledge-learning/africa-capacity-report/africa-capacity-report-acr-2017f (accessed 4 May 2020).

Amorim, R. and Hennig, C. (2015) 'Recovering the number of clusters in data sets with noise features using feature rescaling factors', *Information Sciences*, Vol. 324, pp.126–145.

Bolhaar, J., Gerritsen, S., Kuijpers, S. and van der Wiel, K. (2019) *Experimenting with Dropout Prevention Policies*, CPB discussion paper, CPB Netherlands Bureau for Economic Policy Analysis.

Bouchet, F., Harley, J., Trevors, G. and Azevedo, R. (2013) 'Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning', *Journal of Educational Data Mining*, Vol. 5, No. 1, pp.104–146.

Cherney, K. and Price, M. (2018) *Student Profiling and Negative Implications for Students with Disabilities*, Chapter 6, pp.103–117, John Wiley and Sons, Ltd.

Choi, H.J. and Park, J-H. (2018) 'Testing a path-analytic model of adult dropout in online degree programs', *Computers and Education*, Vol. 116, pp.130–138.

Cortez, P. and Silva, A. (2008) 'Using data mining to predict secondary school student performance', *EUROSIS*.

Duncombe, R. (2016) ‘Mobile phones for agricultural and rural development: a literature review and suggestions for future research’, *The European Journal of Development Research*, Vol. 28, No. 2, pp.213–235.

Dupere, V., Dion, E., Leventhal, T., Archambault, I., Crosnoe, R. and Janosz, M. (2017) ‘High school dropout in proximal context: the triggering role of stressful life events’, *Child Development*, Vol. 89, No. 2, pp.e107–e122.

Fekete (2020) *15 Popup Examples for Effective Customer Feedback Collection* [online] https://www.optimonk.com/15-popup-examples-collection/ (accessed 14 February 2020).

Gower, J.C. and Gower, J.C. (1971) ‘A general coefficient of similarity and some of its properties’, *Biometrics*, December, Vol. 27, No. 4, pp.857–871.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) ‘The WEKA data mining software: an update’, *SIGKDD Explorations*, Vol. 11, No. 1, pp.10–18.

Harrak, F., Bouchet, F., Luengo, V. and Gillois, P. (2018) ‘Profiling students from their questions in a blended learning environment’, *Proceedings of the 8th International Conference on Learning Analytics and Knowledge, LAK ‘18*, pp.102–110, New York, NY, USA. ACM.

Huang, T.H-D. and Kao, H-Y. (2018) *C-3PO: Click-Sequence-Aware Deep Neural Network (DNN)-Based Pop-Ups Recommendation* [online] http://arxiv.org/abs/1803.00458.

Jia, Y., Konold, T. and Cornell, D. (2015) ‘Authoritative school climate and high school dropout rates’, *School Psychology Quarterly*, Vol. 31, No. 2, pp.289–303.

Kotsiantis, S.B. (2014) ‘Bagging and boosting variants for handling classifications problems: a survey’, *The Knowledge Engineering Review*, Vol. 29, No. 1, pp.78–100.

Lee, S. and Chung, J.Y. (2019) ‘The machine learning-based dropout early warning system for improving the performance of dropout prediction’, *Applied Sciences*, Vol. 9, No. 15, p.3093.

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

Lwoga, E.T. and Sangeda, R.Z. (2019) ‘ICTs and development in developing countries: a systematic review of reviews’, *The Electronic Journal of Information Systems in Developing Countries*, Vol. 85, No. 1, p.e12060.

Márquez, C., Cano, A., Romero, C., Mohammad, A., Fardoun, H. and Ventura, S. (2016) ‘Early dropout prediction using data mining: a case study with high school students’, *Expert Systems*, Vol. 33, No. 1, pp.107–124.

President’s Office Regional Administration and Local Government Authority (PORALG) (2016) *Pre-Primary, Primary and Secondary Education Statistics in Brief 2016*, Basic Education Statistics Report.

Quinlan, J.R. (1986) ‘Induction of decision trees’, *Mach. Learn.*, Vol. 1, No. 1, pp.81–106.

Restless Development (2011) *State of the Youth in Tanzania* [online] http://kijanawajibika.com/img/posts/60/59212ba34fb8b.pdf.

Rousseeuw, P. (1987) ‘Silhouettes: a graphical aid to the interpretation and validation of cluster analysis’, *J. Comput. Appl. Math.*, Vol. 20, No. 1, pp.53–65.

Sivakumar, S., Venkataraman, S. and Selvaraj, R. (2016) ‘Predictive modeling of student dropout indicators in educational data mining using improved decision tree’, *Indian Journal of Science and Technology*, Vol. 9, pp.1–5.

Tan, M. and Shao, P. (2015) ‘Prediction of student dropout in e-learning program through the use of machine learning method’, *International Journal of Emerging Technologies in Learning (iJET)*, Vol. 10, No. 1, pp.11–17.

Theunissen, M-J., Bosma, H., Verdonk, P. and Feron, F. (2015) ‘Why wait? early determinants of school dropout in preventive pediatric primary care’, *PLOS ONE*, Vol. 10, No. 11, pp.1–22.

Thornberry, T., Moore, M. and Christenson, R. (1985) ‘The effect of dropping out of high school on subsequent criminal behavior’, *Criminology*, Vol. 23, pp.3–18.

Twaweza (2015) *Uwezo Data on Learning* [online] https://www.twaweza.org/go/uwezo-datasets (accessed 13 February 2020).

UNEconomicCommissionAfrica (2017) *Africa's Youth and Prospects for Inclusive Development 2017* [online] https://www.ohchr.org/Documents/Issues/Youth/UNEconomicCommissionAfrica.pdf (accessed 4 May 2020).

Whitehill, J., Mohan, K., Seaton, D., Rosen, Y. and Tingley, D. (2017) 'MOOC dropout prediction: how to measure accuracy?', *Proceedings of the Fourth Association for Computing Machinery on Learning @ Scale*, pp.161–164 [online] https://scholar.harvard.edu/dtingley/publications/mooc-dropout-prediction-how-measure-accuracy.