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

Neema Mduma [a,*], Khamisi Kalegele [b] and Dina Machuve [c]

[a] *Department of Information and Communication Science and Engineering, The Nelson Mandela African Institution of Science and Technology, NM-AIST, Tanzania*
E-mail: mduman@nm-aist.ac.tz; ORCID: https://orcid.org/0000-0002-4364-3124

[b] *Department of Knowledge Management, Commission for Science and Technology, COSTECH, Tanzania*
E-mail: kalegs03@gmail.com; ORCID: https://orcid.org/0000-0003-3020-6408

[c] *Department of Information and Communication Science and Engineering, The Nelson Mandela African Institution of Science and Technology, NM-AIST, Tanzania*
E-mail: dina.machuve@nm-aist.ac.tz; ORCID: https://orcid.org/0000-0002-8711-5948

Abstract. 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 in 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. To this end, several machine learning algorithms have been proposed in literature. This paper presents a survey of machine learning in education with focus on approaches and techniques for student-dropout prediction. Furthermore, the paper discusses the state of student dropout in developing countries and several performance metrics used by researchers to evaluate machine learning techniques in the context of education with experimental case study. Finally, the paper highlights challenges and future research directions.

Keywords: Machine Learning (ML), Imbalanced learning classification, Secondary education, Evaluation metrics

1. Introduction

Reducing student dropout rates is one of the challenges faced in the education sector globally. This problem brought a major concern in the field of education and policy-making communities [1]. 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 [2]. In Tanzania, for example, student dropout is higher in lower secondary compared to higher level where girls are much less likely than boys to complete secondary education; 30% of girls dropout before reaching form 4 compared to 15% for boys [3]. 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

[*] Corresponding author. E-mail: mduman@nm-aist.ac.tz.

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 [4].

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 [5–7]. Moreover, there exist quite a significant body of literature on machine learning based approaches associated with fighting dropouts [8–10]. The knowledge embodied in literature has the potential to transform the fight against dropout from reactive to proactive. This is a more reality now than ever because the ICTs have already transformed the way we collect and manage data, 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. The study conducted by [11] used the primary data collected in Kenya, although the dataset is not public available. Besides, 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 especially girls. Increasingly, communities of practitioners and researchers are looking at machine learning approaches as a likely solution for achieving dropout-free schools. In this article, a survey of how machine-learning techniques have been used in the fight against dropouts is presented for the purpose of providing 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 paper. The intervention points included issues related to data preprocessing, choosing an algorithm to predict dropouts, and evaluation metrics. In this article, potential machine learning techniques for the three intervention areas are summarized and also the results of demonstrated experiments as part of case study are presented.

1.1. The state of student dropout in developing countries

The issue of student dropout is a serious problem which adversely affects the development of the education sector, this is due to a complex interplay of socio-cultural, economic and structural factors [12]. Schooling, according to the human capital theory, is an investment that generates higher future income for individuals [13]. 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 [14]. However, less attention is paid to improve quality of education to people belongs to any class. In this regard, a [15] report points out, that about one thirty million children in the developing world denied their right to education through dropping out [16].

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) [17]. These were established to focuses 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.

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

In addition to that, gender plays a role on addressing this problem of student dropout. Across the world, females are more likely than males to be out of school, and the poorest girls from the most disadvantaged rural areas tend to have the lowest educational attainment levels. The prevalence of unequal distribution of education in male and female students hinders the development at every stage of a nation. Though, insignificant attention has been dedicated to examining the effects of girl child dropout in schools especially in the developing countries where the problem is widespread, literature has found that girls' dropout rate is significantly higher in rural schools compared to urban schools [2]. Yet, the issue of girl child dropout is a serious problem that dramatically impacts on national development.

Furthermore, [2] observes that though the enrollment in school is almost same for girls and boys, boys have a higher likelihood of continuing school compared to girls. Moreover, girls overall attain less education and tend to drop out earlier as compared to boys. Thus, when dropout rate varies by gender and if girls tend to drop out earlier compared to boys, it manifests that there are some unique factors contributing to the increase in the dropout rate, particularly for girls.

The reasons why females are more likely than males to be out of school relate to social power structures and socially-constructed norms that define the roles that boys and girls should play. These gender roles affect the rights, responsibilities, opportunities and capabilities of males and females, including their access to and treatment in school. Mainly because of gendered perceptions of adolescent girl's roles and responsibilities, in most developing countries, girl's enrollment rates fall when they reach lower secondary school age and then decline further when they reach upper secondary school age [18].

In many developing countries including Tanzania, more than half of the school dropouts are largely attributed to healthy challenges during adolescence. [19], estimated that about 10% of school-age African girls do not attend school during menstruation, or drop out at puberty because of the lack of clean and private sanitation facilities in schools. While the government has to ensure that secondary education remains to be free and compulsory [20], it should also include the commitment to guarantee that intervention programs are informed by the collected statistics so as to contributes in reducing the dropout rates.

1.2. 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 [21]. There are several areas where machine learning can positively impact education. The study conducted by [22], 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 the 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, leveraging the data collected [23].

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 [24, 25], knowledge tracing [26], learning material enhancement [27] and early warning systems [28–30]. The use of these techniques for educational purpose is an promising field aimed at developing methods of exploring data from computational educational settings and discovering meaningful patterns [31].

One of the first applications of machine learning in education has been to help 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 [32], 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 students' 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 Chalapathy Neti shared their 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 which need to be addressed. There is lack of available open-access datasets especially in developing countries; more datasets 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 [9], 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 reveal 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, there is 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.

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 their education, both students and communities lose out on skills, talent and innovation [6]. On addressing student dropout problem, several predictive models were developed 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. Despite the fact that, evaluation of developed predictive models tend to differ but the focus remain 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.

[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

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

2. Approach

During last few years several works 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 useful on understanding the problem and improving measures to address solution.

In this paper we searched the following databases: ResearchGate, Elsevier, 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.

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.

Besides, preliminary survey was conducted to stakeholders and identified that most of the collected data are not in the direct form to support machine learning approach. Most of the data collected are not clean and contains missing values, this is unavoidable problem in dealing with most of the real world data sources. Furthermore, on addressing the problem of student dropout; most of the datasets are facing imbalance problem which needs special attention on developing predictive algorithm. Therefore, this paper provides a suggested approach on addressing the problem of student dropout with focus on three important aspects of data preprocessing, machine learning techniques and evaluation measures to be considered.

3. Material and methods

3.1. Preprocessing of data

Data preprocessing includes data cleaning, normalization, transformation, feature extraction and selection, etc, and the product of data preprocessing is the final training set. In selection, relevant target data is selected from retained data (typically very noisy) and subsequently preprocessed. This goes hand in hand with the integration from multiple sources, filtering irrelevant content and structuring of data according to a target tool [33]. On developing a generalized algorithm, data preprocessing can often have a significant impact. Based on the nature of datasets in many domains, it is well known that data preparation and filtering steps take considerable amount of processing time in ML problems.

Various approaches have been identified in handling missing values, outliers data and numeric values [34]. On addressing the problem of student dropout, one of the common problem which must be considered during preprocessing is data imbalance [35]. Several re-sampling techniques such as under-sampling, over-sampling and hybrids methods can be applied [36].

Under-sampling is a non-heuristic method that aims at creating a subset of the original dataset by eliminating instances until the remaining number of examples is roughly the same as that of the minority class. Over-sampling method create a superset of the original dataset by replicating some instances or creating new instances from existing ones until the number of selected examples plus the original examples of the minority class is roughly equal to that of the majority class. Hybrids method such as SMOTE (Synthetic Minority Oversampling Technique) combines both under-sampling and over-sampling approaches [35].

3.2. 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 [37]. 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 [38].

Several existing works have focused on supervised learning algorithms such as Naive Bayesian Algorithm, Association rules mining, Artificial Neural Network (ANN) based algorithm, Logistic Regression, CART, C4.5, J48, Bayes Net, Simple Logistic, JRip, Random Forest, Logistic Regression analysis, ICRM2 for the classification of the educational dropout student [39]. However, under the classification techniques, Neural Network and Decision Tree are the two methods highly used by the researchers for predicting students performance [40, 41]. The advantage of neural network is that it has the ability to detect all possible interactions between predictors variables [42] and could also do a complete detection without having any doubt even in complex nonlinear relationship between dependent and independent variables [43], while decision tree have been used because of its simplicity and comprehensibility to uncover small or large data structure and predict the value [44].

Unlike supervised, unsupervised learning algorithm is used to identify hidden patterns in unlabeled input data. It refers to provide ability to learn and organise 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 [38].

Several techniques have been proposed on addressing this problem of student dropout using different approaches such as Survival Analysis [10, 45], Matrix Factorization [46–50], and Deep Neural Network [4, 51]. Other approaches such as time series clustering [52, 53] were presented to perform clustering, which are extensively used in recommender systems [54].

On addressing the problem of student dropout, machine learning techniques have been applied in various platforms such as Massive Open On-line Course (MOOC) [4, 55–57] and other Learning Management System (LMS) such as Moodle [48, 52, 58]. These platforms generated datasets which contain information that can be categorized into academic performance, socio-economic and personal information [59]. MOOC platforms such as Coursera and edX is among popular used platforms for student dropout prediction [55]. While, Moodle as a popular Learning Management System [58], provides public datasets such UMN LMS [48]. 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 [52].

3.2.1. Survival Analysis

Survival analysis is used to analyze data in which the time until the event is of interest [60]. It provides various mechanisms to handle such censored data problems that arise in modeling such longitudinal data (also referred to 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 [61].

In the context of education, the use of survival analysis modeling to study student retention was developed. [10] developed a survival analysis framework for early prediction using Cox proportional hazards model (Cox) and applied time-dependent Cox (TD-Cox), which captures time-varying factors and can leverage those information to provide more accurate prediction of student dropout. Certainly, in survival analysis subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs [62]. 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 [63].

3.2.2. Matrix Factorization

Matrix factorization is a clustering machine learning methods that can accommodate framework with some variations [64]. The study presented by [47, 48], described matrix factorization. In this study [48], two classes of methods for building the prediction models were presented. The first class builds these models by using linear regression approaches and the second class builds these models by using 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. One limitation of the standard MF method is that it ignores the sequence in which the students have taken the various courses and as such 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 by [46], proposed a new data transformation model, which is built upon the summarized data matrix of Link-based Cluster Ensembles (LCE). 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.

3.2.3. Deep Neural Network and Probabilistic Graphical Model

Deep neural network (DNN) is an approach based on Artificial Neural Networks (ANN) with multiple hidden layers between the input and output layers [65]. While, Probabilistic Graphical Model (PGM) combine probability theory and graph theory so as to offer a compact graph-based representation of joint probability distributions exploiting conditional independences among the random variables [66]. Similar to shallow ANNs, DNNs can model complex non-linear relationships [67, 68]. Recently, 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.

The study presented by [4] considered two temporal models which are state space models and recurrent neural networks. 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. While it is originated from HMM, it is more general that it can learn to map input sequences to output sequences. Moreover, unlike the standard discrete-state HMM, the state space in described IOHMM formulation is continuous so that the state space can in principle 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 that 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 the influence of a given input on the hidden layer, and therefore on the network output, either decays or blows up exponentially as it 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.

3.3. Evaluation measures for student dropout prediction

Many researchers use various evaluation metrics to measure the performance of student dropout algorithms. On measuring percentage of subjects that are classified correctly, several researchers use accuracy metric [9, 10]. Accuracy is a statistical measure for quantifying the degree of correctness with which a prediction model is able to label the data points [9]. Though accuracy is a widely used metric and is useful in practice, it is also a conservative metric in this context. Further, the metric does not distinguish between the magnitude of errors and it might not be appropriate when the data is imbalance [36, 69].

Classification of imbalanced class size data is where one class is under-represented relative to another [36, 69–76]. According to [77], the imbalanced ratio is about at least 1:10. Since the minority class usually represents the most important concept to be learned, it is difficult to identify it due to exceptional and significant cases [36].

In the context of education, data imbalance is very common classification problem in the field of student retention, mainly because there is large number of students who are registered but there are few number of dropout students [35]. Since accuracy has less effect on minority class than majority class [78], several researchers applied other metrics such as F-measure, Mean Absolute Error (MAE) and Area Under the curve (AUC) on addressing this problem of student dropout.

F-measure is defined as a harmonic mean of precision and recall [10]. Several researchers [79–81] used this metric to evaluate algorithms when predicting student dropout. A high value of F-measure indicates that both precision and recall are reasonably high as defined in equation 1.

$$F_m = \frac{2 \cdot Precison \cdot Recall}{Precision + Recall} \tag{1}$$

where $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$.

From the formula above; TP stand for True Positive, FP for False Positive and FN for False Negative.

The study conducted by [9], presented precision at top K and recall at top K as metrics which are far more informative to educators than traditional precision recall curves. The metrics were used so as to provide precision and recall values of various algorithms at different values of K. Furthermore, the metrics were more informative and help educators to infer the precision and recall of various algorithms at a threshold K of their choice.

Other researchers [9, 10, 48, 80] applied frequently used metric in regression problem such as Mean Absolute Error (MAE) on addressing student dropout, with consideration of time to dropout prediction. MAE is a quantity used to measure how close the predictions are to the actual outcomes as stated in

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{2}$$

where $\hat{y}_i$ is the predicted value and y_i is the true value for subject i.

The limitation of this metric is based on how it treats both underestimating and overestimating of actual value in the same manner. However, in student retention problem, these types of errors have different meaning.

Furthermore, several studies observed AUC on measuring the performance of algorithms used on addressing student dropout problem. AUC is expressed as area under the receiver operating characteristic (ROC) where the curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) under various threshold values [1, 4, 14, 56, 57, 79]. In order to plot a ROC curve, we need to vary the discrimination or classification threshold to generate the corresponding TPR and FPR, so that the AUC measure is invariant to the classification threshold. Numerically speaking, an AUC score is a number in the range [0,1], and the closer the number is to 1, the better the classification performance [4].

Other metrics used to evaluate the performance of models are mean squared error [46, 54], Root-Mean-Square Error (RMSE) [48], error residuals [82], and misclassification rates [52].

4. Experimental framework

4.1. Dataset description

In this section, a through analysis of 13 commonly used machine learning algorithms are presented. The aim of this experiment is to provide data-driven algorithm recommendations to current researchers on the topic. Uwezo data on learning [7] at the country level in Tanzania which was collected in 2015 was used. The dataset consists of 18 features and approximately 61340 samples, which were collected with aim of assessing children's learning levels across hundreds of thousands of households. The dataset were cleaned by replacing the missing values with medians and zeros. Since our target variable is dropout, we checked the distribution of this variable in the dataset and observed that there was imbalance for target variable with only 1.6% dropout as shown in Fig. 1.

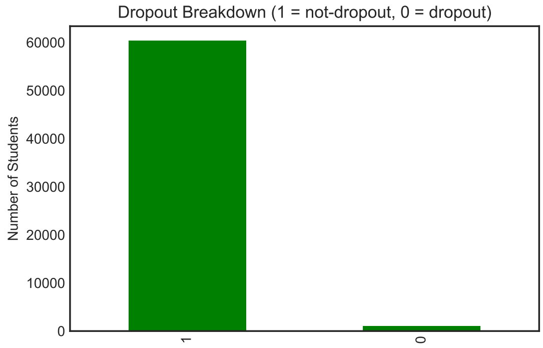

Fig. 1. Dropout distribution training data.

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

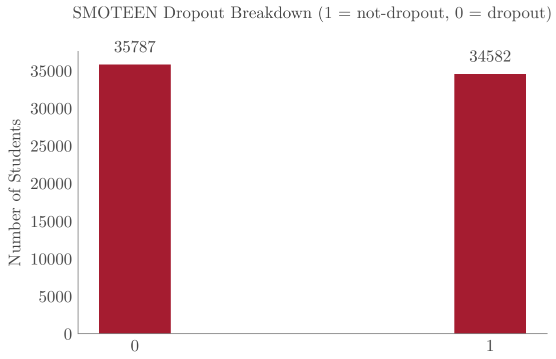

Fig. 2. Dropout-distribution (SMOTEEN).

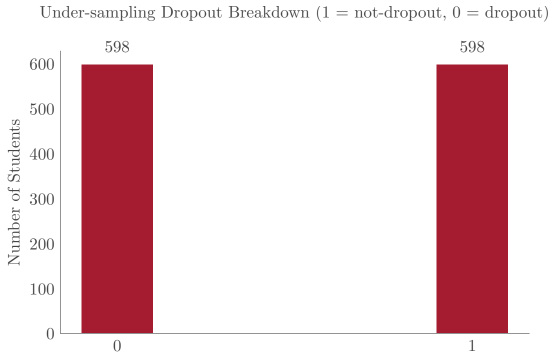

Fig. 3. Dropout-distribution (Under-sampling).

Several approaches such as data re-sampling and generating synthetic samples, just to mention a few, can be used to address this problem. For this problem we opt to use a variety of SMOTE [36]; a popular technique for generating synthetic samples from minority class in order to reinforce its signal. Specifically, SMOTEEN[8] and RandomUnderSampler technique as implemented in Imbalanced-Learn[9] were used. SMOTEEN combine over- and under-sampling using SMOTE and Edited Nearest Neighbour (EN) to generate more minority class where RandomUnderSampler is a fast and easy way to balance the minority class by randomly selecting a subset of data for the targeted classes as observed in Fig. 2, we also observed dropout distribution (Under-sampling) as shown in Fig. 3. The dataset was separated into training (60%), test (20%) and validation (20%).

4.2. Experimental procedures

In order to conduct this experiment, the dataset was divided into three partitions of training (60%), test (20%) and validation (20%). The sampling technique was applied only to the training set and the first experiment of building the model was conducted. Thereafter, the second experiment was to tune the

[8] combine over- and under-sampling using SMOTE and Edited Nearest Neighbor (ENN)

[9] A Python library containing various algorithms to handle imbalanced data sets as well as producing imbalanced data sets:http://contrib.scikit-learn.org/imbalanced-learn/stable/index.html

best performed models in order to improve their predictive performance. This experiment was followed by combining train and validation sets in order to generate a big training set and applied the sampling technique to the new training set. We then evaluate the model using unseen test set in order to observe how the model will behave in the real environment which is imbalance. The overall experiment procedure is summarized in Fig. 4 and stratified k-fold cross validation was used in each experiment in order to avoid over fitting. We use $k = 5$ fold out-of-bag overall cross validation instead of averaging over folds. The entire process involves executing all selected classification algorithms in which all executions are repeated 5 times using training (60%), test (20%) and validation (20%) partitions of the data set. This cross-validation procedure divides the data set into 5 roughly equal parts. For each part, it trains the model using the four remaining parts and computes the test error by classifying the given part. Finally, the results for the five test partitions were averaged.

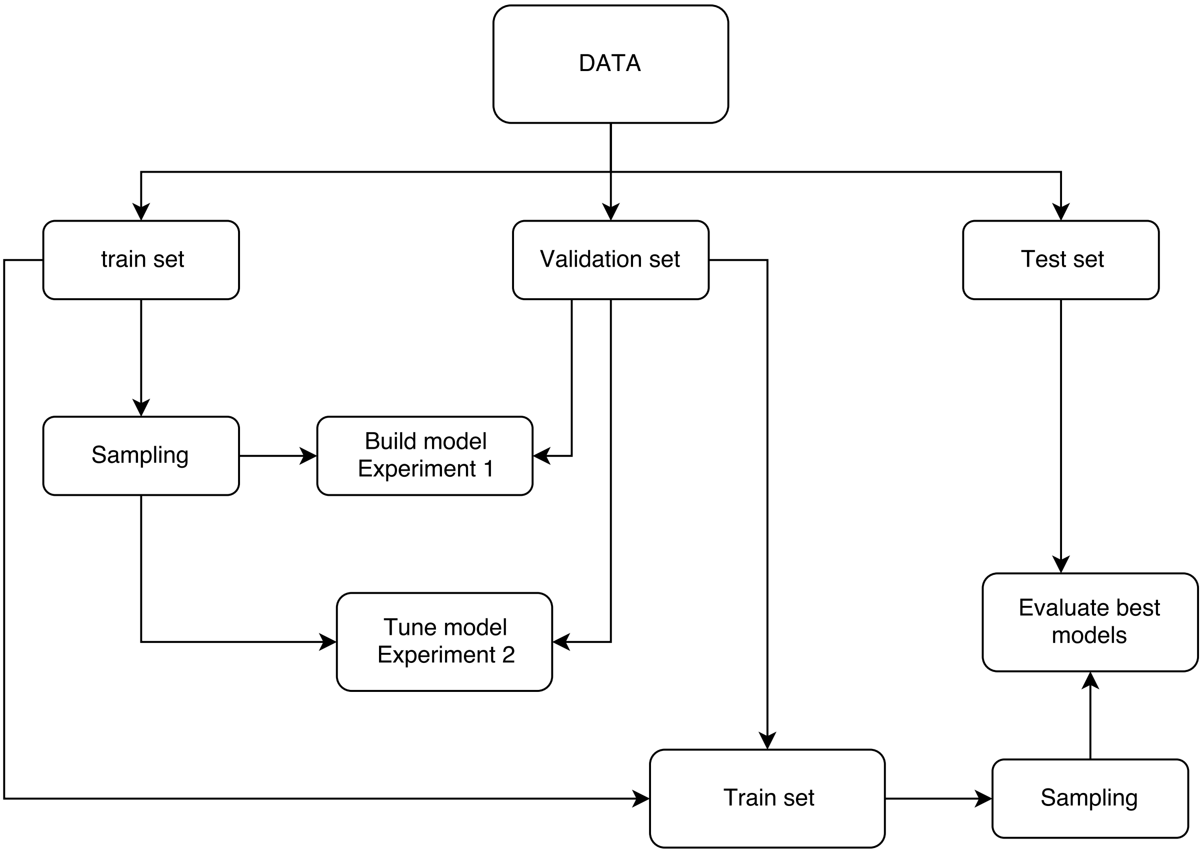

Fig. 4. Experiment procedure.

4.3. Evaluation metrics

To evaluate the model, Geometric Mean (G_m), F-measure (F_m) and Adjusted Geometric Mean (AG_m) metrics were used. The choice of these metrics is attributed to the fact that in imbalanced domains, the evaluation of the classifiers' performance must be carried out using specific metrics in order to take into account the class distribution [36].

Therefore, the G_m is a measure of the ability of a classifier to balance TPrate (sensitivity) and TNrate (specificity) [69, 83] as defined in equation 3. This measure is maximum when TPrate and TNrate are equal. Furthermore, in order to ensure TPrate to the changes in the positive predictive value (precision) than in TPrate, F_m is used as defined in equation 1. Besides, AGm as defined in equation 4 was used to obtain the highest TPrate without decreasing too much the TNrate [36].

$$G_m = \sqrt{(\text{TPrate} \cdot \text{TNrate})} \tag{3}$$

$$AG_m = \begin{cases} \frac{\text{GM}+\text{TNrate}\cdot(FP+TN)}{1+FP+TN} & \text{if TPrate} > 0, \\ 0 & \text{if TPrate} = 0 \end{cases} \tag{4}$$

where:

- TN is true negative, TP is true positive, FN is false negative and FP is false positive.
- TPrate $= \frac{\text{TP}}{\text{TP+FN}}$ the percentage of positive instances correctly classified.
- TNrate $= \frac{\text{TN}}{\text{FP+TN}}$ the percentage of negative instances correctly classified.

4.4. Experiment 1: Model selection

The aim of this experiment is to identify classifier with the best performance for this problem. In this phase, selection of classifiers were based in all domains including linear, ensemble and neural networks classifiers with consideration of classification and nature of the dataset. We selected 13 mostly used classifiers: K-Nearest-Neighbors (KNN), Gaussian Naive Bayes (GNB), Logistic Regression classifier (LR), Linear Discriminant Analysis (LDA), Decision Tree (DTree), Random Forest (RForest), Adaptive Boosting (AdaBoost), Multilayer perceptron (MLP), SGD Classifier which is regularized linear models with Stochastic Gradient Descent (SGD) learning, Extra Tree classifier (EXT), Gradient Boosting Classifier (GBC), Bernoulli Naive Bayes (BNB) and Quadratic Discriminant Analysis (QDA). The experiment is repeated for three different cases: when no sampling is used, when under-sampling is used and when over sampling (SMOTEEN) is used. The experimental results for both cases are presented in Fig. 7, 10 and 13.

To select the best classifiers, we only consider validation results because it give an estimate on how the classifier will perform on actual dataset which is imbalance. From the result presented in Fig. 7 three classifiers: LDA, LR and MLP show better generalization results. They show better validation result for the three metrics used. Considering the case when under-sampling is used, Fig. 10 all classifiers have considerably the same generalization results for both metrics with exception to BNB, SGD, and MLP which show lower G_m. The experiment conducted without sampling revel that, only LR classifier show better performance than others. However, the score rates is less than 1 for AG_m as compared to when LR is used with oversampling case. Therefore, for the next experiment we only consider the following three classifiers: LR, LDA and MLP with oversampling case.

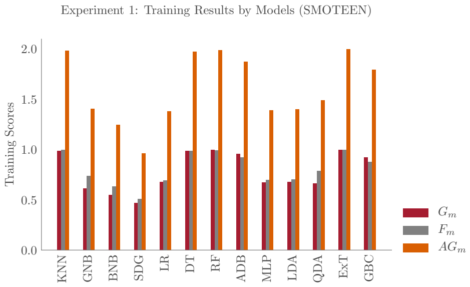

Fig. 5. Experiment 1: Training results (over-sampling).

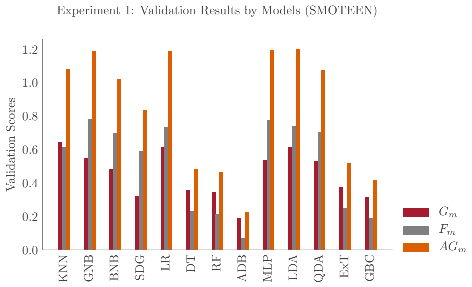

Fig. 6. Experiment 1: Validation results (oversampling).

Fig. 7. Experiment 1: Over sampling.

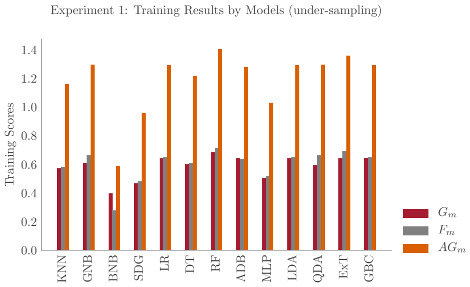

Fig. 8. Experiment 1: Training results (under-sampling).

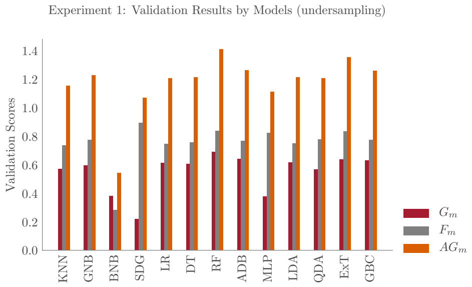

Fig. 9. Experiment 1: Validation results (under--sampling).

Fig. 10. Experiment 1: Under sampling.

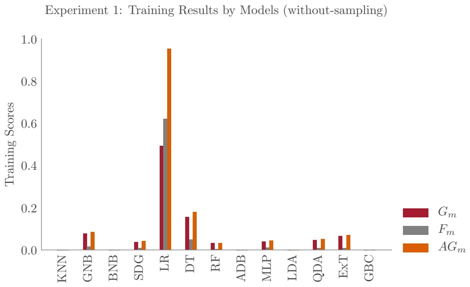

Fig. 11. Experiment 1: Training results (no-sampling).

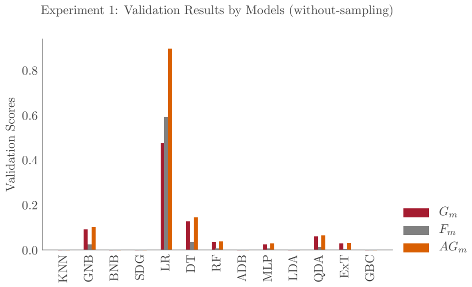

Fig. 12. Experiment 1: Validation results (no-sampling).

Fig. 13. Experiment 1: without sampling.

4.5. Experiment 2: Hyper-parameter optimization

This experiment aim to show the significance of hyper parameter tuning on improving predictive performance. Most ML algorithms contain several hyper parameters that can affect performance significantly (for example, the number hidden layers in MLP classifier). In this experiment three selected classifiers were tuned: LR, LDA and MLP to further improve their performance. We employed hyper-parameter tuning via cross-validation and identified the best parameters for each classifier as presented in Table 1. We then evaluated the models by comparing their validation results as shown in Table 2.

Table 1
Model parameter

Classifier	**Parameter**
LR	fit_intercept:True, tol:1, C:0.001, Penalty:'l1'
LDA	shrinkager:'auto, tol:1e-06, solver:'lsqr'
MLP	solver:'adam', learning_rate_int:0.001, shuffle:True, hidden_layer_size:10, alpha:1, early_stoping: True

The experimental results allow us to measure the extent to which hyper parameter tuning improves each algorithm's performance compared to its baseline settings.

To further improve the performance of the models we employed ensemble technique. Ensemble method is one of the popular approach for improving machine learning algorithms. This approach create multiple models and then combine them to produce improved results. Several ensemble techniques such as bagging, boosting and voting have been extensively use in the literature [84]. For this problem, voting ensemble technique was appropriate. We employed voting (stacking) by soft combined the three tuned classifiers LR2, LDA2 and MLP2. The tuned classifiers where then trained on the new training set obtained by combining validation and training set used in previous experiment. To evaluate the generalization performance, the models were tested on unseen tested data. The result for this experiment is presented in Table 2.

Table 2
Experiment 2: Results

		LR	LR2	MLP	MLP2	LDA	LDA2	ENB
Validation Scores	G_m	0.616	0.617	0.568	0.606	0.614	0.613	**0.623**
	AG_m	1.191	**1.265**	1.110	1.225	1.199	1.198	1.262
	F_m	0.732	**0.787**	0.683	0.766	0.741	0.740	0.781
Test Scores	G_m	0.610	0.612	0.513	0.614	0.612	0.612	**0.635**
	AG_m	1.183	1.260	1.336	1.167	1.200	1.200	**1.277**
	F_m	0.731	**0.788**	0.664	0.714	0.744	0.744	0.784

From Table 2, it can be seen that the stacking classifier (ENB) show considerably better validation and test results followed by the tuned logistic regression model (LR2).

4.6. Experiment 3: Feature Importance

The experiment aimed at identifying the contribution of each features on the prediction performance by automatically selecting features that are most relevant to the dropout predictive modeling. Our dataset consists of 18 features as described in Table 3. This experiment was accomplished by directly measuring the impact of each feature on the model performance (G_m) obtained by permuting the values of each feature and measuring how much the permutation decreases the model performance. Thus for unimportant features, the permutation will have little or no effect on model accuracy, while permuting important variables should significantly decrease it. The results presented in Fig. 14, show clearly that student sex have strong contribution on the dropout prediction performance.

Table 3
Summary of the features for the dataset

No.	Feature description	Type of data
1.	Main source of household income (Income)	Multi nominal
2.	Boy's Pupil Latrines Ratio (BPLR)	Number
3.	School has girl's privacy room (SGR)	Binary nominal
4.	Region	Nominal
5.	District	Nominal
6.	Village	Nominal
7.	Student gender (Sex)	Binary nominal
8.	Parent who check his/her child's exercise book once in a week (PCCB)	Binary nominal
9.	Household meals per day (MLPD)	Multi nominal
10.	Student who did read any book with his/her parent in last week (SPB)	Binary nominal
11.	Parent who discuss his/her child's progress with teacher last term (PTD)	Binary nominal
12.	Student age (Age)	Number
13.	Enumeration Area type (EAarea)	Multi nominal
14.	Household size (HHsize)	Number
15.	Girl's Pupil Latrines Ratio (GPLR)	Number
16.	Parent Teacher Meeting Ratio (PTMR)	Number
17.	Pupil Classroom Ratio (PCR)	Number
18.	Pupil Teacher Ratio (PTR)	Number

5. Discussion and conclusions

5.1. Open challenge and future research direction

In the previous sections we have presented a survey 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 problem of student dropout. The limitation of public datasets from developing countries [79], brought need to develop more datasets from different geographical location. However, cost and time must be acquired to

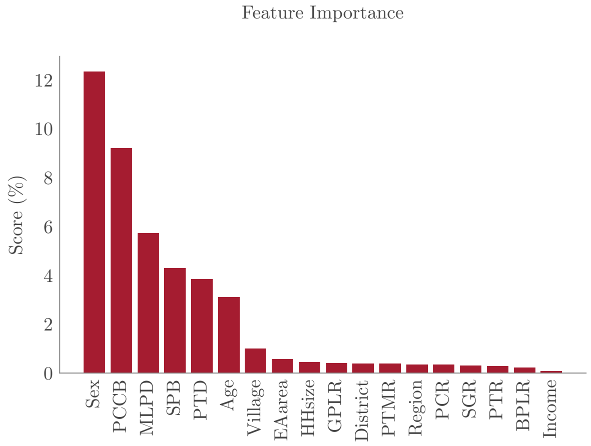

Fig. 14. Feature selection.

accommodate data collection process. Furthermore, to the knowledge of researchers, there are only few research 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.

Second, most of the presented works have focused on providing early prediction only [9]. Therefore, future work should focus on facilitating a more robust and comprehensive early warning systems for students dropout which can identify 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.

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 [35], 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 the problem due to the issue of limited resources which face many school districts [9]. This will involve the use of new sources school level data and applying additional machine learning approaches to improve predictive power of the proposed algorithm. The algorithm will enable relevant authorities to effectively and accurately plan, formulate policies, and make decisions on measures to address the problem.

Fifth, many researchers tackling imbalanced data using commonly approaches as presented in the experimental case study. Further research should focus on the structure and nature of examples in minority classes in order to gain a better insight into the source of learning difficulties [85]. Furthermore, the experimental results presented in this paper as part of case study support other researchers' finding on dropout rate with gender association [2]. However, future research should be conducted on addressing the problem of student dropout with consideration of many datasets and time efficiency so as to provide an effective student dropout model.

5.2. Conclusions

In this work, 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 on 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 research focus on providing early prediction rather than including ranking and forecasting mechanisms on addressing the problem of student dropout.

Fourth, 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.

Fifth, the experimental case study empirically assessed 13 supervised classification algorithms on a set of approximately 61340 supervised classification dataset in order to provide a contemporary set of recommendations to researchers who wish to apply machine learning algorithms to their data with consideration of data imbalanced problem. The three classifiers LR, LDA and MLP have proven superior to all the other classifiers by achieving highest performance metrics when over-sampling technique is employed. Furthermore, we show that hyper parameter tuning improves each algorithm's performance compared to its baseline settings and stacking these classifiers improve the overall predictive performance. We also show the contribution of each features on prediction performance with student sex being the leading feature.

Acknowledgments

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

References

[1] L. Aulck, N. Velagapudi, J. Blumenstock and J. West, Predicting Student Dropout in Higher Education (2016). http://arxiv.org/abs/1606.06364.

[2] S.M. Shahidul and A.H.M.Z. Karim, Factors contributing to school dropout among the girls: a review of literature **3**(2) (2015), 25–36.

[3] BEST, Pre-Primary, Primary and Secondary Education Statistics in Brief 2016 The United Republic of Tanzania President's Office Regional Administration and Local Government (2015).

[4] M. Fei and D.-Y. Yeung, Temporal Models for Predicting Student Dropout in Massive Open Online Courses, *2015 IEEE International Conference on Data Mining Workshop (ICDMW)* (2015), 256–263. ISBN 978-1-4673-8493-3. doi:10.1109/ICDMW.2015.174. http://ieeexplore.ieee.org/document/7395679/.

[5] K. Adhatrao, A. Gaykar, A. Dhawan, R. Jha and V. Honrao, Predicting Students' Performance Using Id3 and C4.5 Classification Algorithms, *International Journal of Data Mining and Knowledge Management Process* **3**(5) (2013), 39–52, ISSN 2231007X. doi:10.5121/ijdkp.2013.3504.

[6] M. Durairaj and C. Vijitha, Educational data mining for prediction of student performance using clustering algorithms, *International Journal of Computer Science and Information Technologies (IJCSIT)* **5**(4) (2014), 5987–5991, ISSN 0975-9646.

[7] J.F. Chen, H.N. Hsieh and Q.H. Do, Predicting student academic performance: A comparison of two meta-heuristic algorithms inspired by cuckoo birds for training neural networks, *Algorithms* **7**(4) (2014), 538–553, ISSN 19994893. doi:10.3390/a7040538.

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

[9] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani and K.L. Addison, A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes, *Kdd* (2015), 1909–1918. ISBN 9781450336642. doi:10.1145/2783258.2788620.

[10] S. Ameri, M.J. Fard, R.B. Chinnam and C.K. Reddy, Survival Analysis based Framework for Early Prediction of Student Dropouts (2016). ISBN 9781450340731. doi:10.1145/2983323.2983351.

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

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

[13] R. Patron, Early school dropouts in developing countries: An equity issue? The Uruguayan case, *University of Uruguay* (2014), 13. http://www.ecineq.org/ecineq{_}ba/papers/Patron.pdf.

[14] R. Halland, C. Igel and S. Alstrup, 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* (2015), 22–24. ISBN 9782875870148.

[15] UNESCO, UNESCO Global Partnership for Girls' and Women's Education- One Year On (2011). http://www.unesco.org/eri/cp/factsheets{_}ed/TZ{_}EDFactSheet.pdf.

[16] L.A. Choudhary AI, Economic Effects of Student Dropouts: A Comparative Study, *Journal of Global Economics* **03**(02) (2015), 2–5, ISSN 23754389. doi:10.4172/2375-4389.1000137. http://www.esciencecentral.org/journals/economic-effects-of-student-dropouts-a-comparative-study-2375-4389-1000137.php?aid=57059.

[17] TAMISEMI, The United Republic of Tanzania Ministry of Education and Culture (2004), 2004–2009.

[18] G. Subrahmanyam, Gender perspectives on causes and effects of school dropouts (2016).

[19] UNICEF, UNICEF Water, Sanitation and Hygiene Annual Report (2006).

[20] Human RIghts Watch, *I Had a Dream to Finish SchoolĂİ:Barriers to Secondary Education in Tanzania*, 2017, pp. 1–108. ISBN 9781623134419.

[21] M.I. Jordan and T.M. Mitchell, Machine learning: Trends, perspectives, and prospects, *Science* **349**(6245) (2015), 255–260, ISSN 10959203. ISBN 0036-8075, 0036-8075. doi:10.1126/science.aaa8415.

[22] Center for Digital Technology and Management, *THE FUTURE OF EDUCATION TREND REPORT 2015*, 2015. ISBN 9783981553871.

[23] K. Lee, Large-Scale and Interpretable Collaborative Filtering for Educational Data (2017), 1–7.

[24] A.S. Lan, C. Studer and R.G. Baraniuk, Time-varying Learning and Content Analytics via Sparse Factor Analysis (2014). ISBN 9781450329569.

[25] A.E. Waters, C. Studer and R.G. Baraniuk, Sparse Factor Analysis for Learning and Content Analytics **15** (2014), 1959–2008.

[26] M.V. Yudelson, K.R. Koedinger and G.J. Gordon, Individualized Bayesian Knowledge Tracing Models (2013), 1–10.

[27] R. Agrawal, Mining Videos from the Web for Electronic Textbooks (2014).

[28] H.P. Beck and W.D. Davidson, Establishing an Early Warning System : Predicting Low Grades in College Students from Survey of Academic Orientations ... (2016). doi:10.1023/A.

[29] A. Brundage, The use of early warning systems to promote success for all students (2014). http://www.fldoe.org/core/fileparse.php/5423/urlt/ews.pdf.

[30] US Department of Education, Definition of Early Warning Systems Research on Early Warning Systems Issue Brief: Early Warning Systems (2016), 1–13. http://ies.ed.gov/ncee/edlabs/projects/ews.asp..

[31] S. Nunn, J.T. Avella, T. Kanai and M. Kebritchi, Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review, *Online Learning* **20**(2) (2016), 13–29, ISSN 2472-5730. ISBN 1939-5256. doi:10.24059/olj.v20i2.790. https://olj.onlinelearningconsortium.org/index.php/olj/article/view/790.

[32] S.B. Kotsiantis, Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades, *Artificial Intelligence Review* **37**(4) (2012), 331–344, ISSN 02692821. ISBN 0269-2821\r1573-7462. doi:10.1007/s10462-011-9234-x.

[33] K. Kalegele, K. Sasai, H. Takahashi, G. Kitagata and T. Kinoshita, Four Decades of Data Mining in Network and Systems Management, *IEEE Transactions on Knowledge and Data Engineering* **27**(10) (2015), 2700–2716, ISSN 10414347. doi:10.1109/TKDE.2015.2426713.

[34] S. Shahul, S. Suneel, M.A. Rahaman and &. Swathi, A Study of Data Pre-Processing Techniques for Machine Learning Algorithm to Predict Software Effort Estimation, *Imperial Journal of Interdisciplinary Research* **2**(6) (2016), 2454–1362. http://www.onlinejournal.in.

[35] D. Thammasiri, D. Delen, P. Meesad and N. Kasap, A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition, *Expert Systems with Applications* **41**(2) (2014), 321–330, ISSN 09574174. ISBN 0957-4174. doi:10.1016/j.eswa.2013.07.046. http://dx.doi.org/10.1016/j.eswa.2013.07.046.

[36] V. López, A. Fernández, S. García, V. Palade and F. Herrera, An insight into classification with imbalanced data : Empirical results and current trends on using data intrinsic characteristics, *Information Sciences* **250** (2013), 113–141, ISSN 0020-0255. doi:10.1016/j.ins.2013.07.007. http://dx.doi.org/10.1016/j.ins.2013.07.007.

[37] Erik G., Introduction to Supervised Learning (2014), 1–5. https://people.cs.umass.edu/{~}elm/Teaching/Docs/supervised2014a.pdf{%}0Ahttp://people.cs.umass.edu/{~}elm/Teaching/Docs/supervised2014a.pdf.

[38] R. Sathya and A. Abraham, Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification, *International Journal of Advanced Research in Artificial Intelligence* **2**(2) (2013), 34–38, ISSN 21654069. ISBN 9781479923410. doi:10.14569/IJARAI.2013.020206. http://thesai.org/Publications/ViewPaper?Volume=2{&}Issue=2{&}Code=IJARAI{&}SerialNo=6.

[39] M. Kumar, A.J. Singh and D. Handa, Literature Survey on Educational Dropout Prediction, *I.J. Education and Management Engineering* **2**(March) (2017), 8–19, ISSN 23053623. doi:10.5815/ijeme.2017.02.02.

[40] A.M. Shahiri, W. Husain and N.A. Rashid, A Review on Predicting Student's Performance Using Data Mining Techniques, *Procedia Computer Science* **72** (2015), 414–422, ISSN 18770509. ISBN 18770509 (ISSN). doi:10.1016/j.procs.2015.12.157. http://dx.doi.org/10.1016/j.procs.2015.12.157.

[41] H.R. Joseph, 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* (2014), 251–254. ISBN 9781479940974. doi:10.1109/GHTC-SAS.2014.6967592.

[42] G. Gray, C. McGuinness and P. Owende, An application of classification models to predict learner progression in tertiary education, *2014 4th IEEE International Advance Computing Conference, IACC 2014* (2014), 549–554. ISBN 978-1-4799-2572-8. doi:10.1109/IAdCC.2014.6779384.

[43] J.-l.A. Arsad, Pauziah Mohd Buniyamin, Norlida Manan, A neural network students' performance prediction model (NNSPPM), *Smart Instrumentation, Measurement and Applications (ICSIMA), 2013 IEEE International Conference on* (2013), 1–5. ISBN 9781479908431. doi:10.1109/ICSIMA.2013.6717966.

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

[45] S. Ameri, Survival Analysis Approach For Early Prediction Of Student Dropout (2015).

[46] N. Iam-On and T. Boongoen, Generating descriptive model for student dropout: a review of clustering approach, *Human-centric Computing and Information Sciences* **7**(1) (2017), 1, ISSN 2192-1962. ISBN 2192-1962. doi:10.1186/s13673-016-0083-0. http://hcis-journal.springeropen.com/articles/10.1186/s13673-016-0083-0.

[47] Q. Hu and H. Rangwala, Enriching Course-Specific Regression Models with Content Features for Grade Prediction (2016). ISBN 1234567245.

[48] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis and H. Rangwala, –okay–Predicting Student Performance Using Personalized Analytics, *Computer* **49**(4) (2016), 61–69, ISSN 00189162. ISBN 0018-9162 VO - 49. doi:10.1109/MC.2016.119.

[49] Z. Iqbal, J. Qadir, A.N. Mian and F. Kamiran, Machine Learning Based Student Grade Prediction: A Case Study (2017), 1–22. https://arxiv.org/pdf/1708.08744.pdf.

[50] A.R. Babu, 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) (2015), 251–262, ISSN 0974-5645. doi:10.17485/ijst/2015/v8iS7/.

[51] W. Wang, H. Yu and C. Miao, Deep Model for Dropout Prediction in MOOCs, *Proceedings of the 2nd International Conference on Crowd Science and Engineering - ICCSE'17* (2017), 26–32. ISBN 9781450353755. doi:10.1145/3126973.3126990. http://dl.acm.org/citation.cfm?doid=3126973.3126990.

[52] J.L. Hung, M.C. Wang, S. Wang, M. Abdelrasoul, Y. Li and W. He, Identifying At-Risk Students for Early Interventions - A Time-Series Clustering Approach, *IEEE Transactions on Emerging Topics in Computing* **5**(1) (2017), 45–55, ISSN 21686750. ISBN 2168-6750 VO - 5. doi:10.1109/TETC.2015.2504239.

[53] E. Młynarska, D. Greene and P. Cunningham, Time series clustering of Moodle activity data, *CEUR Workshop Proceedings* **1751** (2016), 104–115, ISSN 16130073.

[54] J. Xu, K.H. Moon and M. van der Schaar, A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs, *IEEE Journal of Selected Topics in Signal Processing* **11**(5) (2017), 742–753, ISSN 1932-4553. doi:10.1109/JSTSP.2017.2692560. http://ieeexplore.ieee.org/document/7894238/.

[55] Y. Chen, Q. Chen, M. Zhao, S. Boyer, K. Veeramachaneni and H. Qu, 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* (2017), 111–120. ISBN 9781509056613. doi:10.1109/VAST.2016.7883517.

[56] J. Liang, C. Li and L. Zheng, Machine learning application in MOOCs: Dropout prediction, *ICCSE 2016 - 11th International Conference on Computer Science and Education* (2016), 52–57. ISBN 9781509022182. doi:10.1109/ICCSE.2016.7581554.

[57] L.P. Prieto, M.J. Rodríguez-Triana, M. Kusmin and M. Laanpere, Smart school multimodal dataset and challenges, *CEUR Workshop Proceedings* **1828** (2017), 53–59, ISSN 16130073. ISBN 9781450321389. doi:10.1145/1235.

[58] M.A. Santana, E.B. Costa, B.F.S. Neto, I.C.L. Silva and J.B.A. Rego, A predictive model for identifying students with dropout profiles in online courses, *CEUR Workshop Proceedings* **1446** (2015), ISSN 16130073.

[59] C. Lei and K.F. Li, Academic Performance Predictors, in: *Proceedings - IEEE 29th International Conference on Advanced Information Networking and Applications Workshops, WAINA 2015*, 2015. ISBN 9781479917747. doi:10.1109/WAINA.2015.114.

[60] O.O. Kartal, USING SURVIVAL ANALYSIS TO INVESTIGATE THE PERSISTENCE OF STUDENTS IN AN INTRODUCTORY INFORMATION TECHNOLOGY COURSE AT METU, PhD thesis, 2015.

[61] P. Wang, Y. Li and C.K. Reddy, Machine Learning for Survival Analysis: A Survey, *ACM Comput. Surv. Article* **1**(1) (2017), 38. doi:0000001.0000001. http://dmkd.cs.vt.edu/papers/CSUR17.pdf.

[62] Y. Li, J. Wang, J. Ye and C.K. Reddy, A Multi-Task Learning Formulation for Survival Analysis, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (2016), 1715–1724. ISBN 9781450342322. doi:10.1145/2939672.2939857. http://dl.acm.org/citation.cfm?doid=2939672.2939857.

[63] M.J. Bani and M. Haji, College Student Retention: When Do We Losing Them? (2017). ISBN 9789881404756. http://arxiv.org/abs/1707.06210.

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

[65] L. Deng and D. Yu, Deep Learning: Methods and Applications, *Foundations and Trends® in Signal Processing* **7**(3–4) (2014), 197–387, ISSN 1932-8346. ISBN 9781405161251. doi:10.1561/2000000039. http://nowpublishers.com/articles/foundations-and-trends-in-signal-processing/SIG-039.

[66] F. Pernkopf, R. Peharz and S. Tschiatschek, *Introduction to Probabilistic Graphical Models Introduction*, 2013, pp. 1–60.

[67] S. Mun, M. Shin, S. Shon, W. Kim, D.K. Han and H. Ko, DNN transfer learning based non-linear feature extraction for acoustic event classification, *IEICE Transactions on Information and Systems* **E100D**(9) (2017), 1–4, ISSN 17451361. doi:10.1587/transinf.2017EDL8048.

[68] V. Ramachandra and K. Way, Deep Learning for Causal Inference (2018).

[69] W.J. Lin and J.J. Chen, Class-imbalanced classifiers for high-dimensional data, *Briefings in Bioinformatics* **14**(1) (2013), 13–26, ISSN 14675463. ISBN 1477-4054. doi:10.1093/bib/bbs006.

[70] B. Krawczyk and G.S. B, Pattern Recognition and Machine Intelligence **9124** (2015), 535–544. ISBN 978-3-319-19940-5. doi:10.1007/978-3-319-19941-2. http://link.springer.com/10.1007/978-3-319-19941-2.

[71] M. Galar, A. Fernández, E. Barrenechea, H. Bustince and F. Herrera, New Ordering-Based Pruning Metrics for Ensembles of Classifiers in Imbalanced Datasets (2016). ISBN 9783319262277. doi:10.1007/978-3-319-26227-7.

[72] B. Krawczyk, Combining One-vs-One Decomposition and Ensemble Learning for Multi-class (2015), 27–36. ISBN 9783319262277. doi:10.1007/978-3-319-26227-7.

[73] K. Borowska and M. Topczewska, New Data Level Approach for Imbalanced Data Classification Improvement (2016), 283–294. ISBN 9783319262277. doi:10.1007/978-3-319-26227-7.

[74] J. Stefanowski, On Properties of Undersampling Bagging (2016). ISBN 9783319262277. doi:10.1007/978-3-319-26227-7.

[75] R.U. Mazumder, S.A. Begum and D. Biswas, Rough Fuzzy Classi fi cation for Class Imbalanced Data (2015). ISBN 9788132222170. doi:10.1007/978-81-322-2217-0.

[76] L. Abdi and S. Hashemi, An Ensemble Pruning Approach Based on Reinforcement Learning in Presence of Multi-class Imbalanced Data (2014). ISBN 9788132217718. doi:10.1007/978-81-322-1771-8.

[77] T. Gao, Hybrid classification approach of SMOTE and instance selection for imbalanced datasets (2015).

[78] R. Longadge, S.S. Dongre and L. Malik, Class imbalance problem in data mining: review, *International Journal of Computer Science and Network* **2**(1) (2013), 83–87, ISSN 2277-5420. ISBN 978-1-4673-5563-6. doi:10.1109/SIU.2013.6531574.

[79] M. Mgala and A. Mbogho, Data-driven Intervention-level Prediction Modeling for Academic Performance, *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development* (2015), 2–128. ISBN 978-1-4503-3163-0. doi:10.1145/2737856.2738012. http://doi.acm.org/10.1145/2737856.2738012.

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

[81] L. Aulck, R. Aras, L. Li, C.L. Heureux, P. Lu and J. West, STEM-ming the Tide : Predicting STEM attrition using student transcript data (2017). ISBN 1234567245.

[82] N. Poh and I. Smythe, 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* (2015), 416–421. ISBN 9781479945191. doi:10.1109/CIDM.2014.7008698.

[83] C. Márquez-Vera, A. Cano, C. Romero, A.Y.M. Noaman, H. Mousa Fardoun and S. Ventura, Early dropout prediction using data mining: A case study with high school students, *Expert Systems* **33**(1) (2016), 107–124, ISSN 14680394. ISBN 1468-0394. doi:10.1111/exsy.12135.

[84] P.T. Dalvi and N. Vernekar, Anemia Detection using Ensemble Learning Techniques and Statistical Models (2016), 1747–1751. ISBN 9781509007745.

[85] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, *Progress in Artificial Intelligence* **5**(4) (2016), 221–232, ISSN 2192-6352. doi:10.1007/s13748-016-0094-0. http://link.springer.com/10.1007/s13748-016-0094-0.