2019-05-06

Machine learning approach for reducing students dropout rates

Mduma, Neema

International Journal of Advanced Computer Research

http://dx.doi.org/10.19101/IJACR.2018.839045

Research Article

International Journal of Advanced Computer Research, Vol 9(42)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2018.839045

Machine learning approach for reducing students dropout rates

Neema Mduma[1*], Khamisi Kalegele[2] and Dina Machuve[3]
Research Scholar, The Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania[1]
Lecturer and Researcher, The Tanzania Commission of Science and Technology, Arusha, Tanzania[2]
Lecturer and Researcher, The Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania[3]

Received: 29-November-2018; Revised: 31-January-2019; Accepted: 05-February-2019

Abstract
School dropout is a widely recognized serious issue in developing countries. On the other hand, machine learning techniques have gained much attention on addressing this problem. This paper, presents a thorough analysis of four supervised learning classifiers that represent linear, ensemble, instance and neural networks on Uwezo Annual Learning Assessment datasets for Tanzania as a case study. The goal of the study is to provide data-driven algorithm recommendations to current researchers on the topic. Using three metrics: geometric mean, F-measure and adjusted geometric mean, we assessed and quantified the effect of different sampling techniques on the imbalanced dataset for model selection. We further indicate the significance of hyper parameter tuning in improving predictive performance. The results indicate that two classifiers: logistic regression and multilayer perceptron achieve the highest performance when over-sampling technique was employed. Furthermore, hyper parameter tuning improves each algorithm's performance compared to its baseline settings and stacking these classifiers improves the overall predictive performance.

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

1.Introduction
Reducing student dropout rates are one of the challenges faced by many school districts globally. A growing body of literature indicates high rates of students drop-out of school, especially in the developing world [1]. In addressing this problem, machine learning techniques have gained much attention in recent years [2−4]. This is attributed by the fact that machine learning provides a convenient way to solve student dropout problems and delivering good guarantees for the solutions [5, 6]. To this end, a substantial amount of literature that focuses on predicting student dropout has been presented [7−9]. Different machine learning techniques such as decision tree [2]; artificial neural networks, matrix factorization [3, 5, 10−12], Deep neural network [13], probabilistic graphical models [14, 15] and survival analysis [7] have been applied to develop predictive algorithms.

Other approaches such as time series clustering [16, 17] were presented to perform clustering, which are extensively used in recommender systems [3].

*Author for correspondence

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 in addressing this problem. The study conducted by Mgala, used the primary data collected from primary schools in Kenya, although the dataset is not publicly available [1]. The limitation of public datasets from developing countries brought the need to develop more datasets from different geographical locations. However, cost and time must be acquired to accommodate the data collection process. Besides, to the knowledge of researchers, there are only few studies which have 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.

Machine learning techniques have been applied in various platforms such as a massive open on-line course (MOOC). MOOC platforms such as Coursera and the edX is among popular used platforms for student dropout prediction [9] and other learning

management system (LMS) such as Moodle [16]. On addressing the problem of student dropout, several existing works have focused on supervised learning algorithms such as a naive Bayesian algorithm, association rules mining, artificial neural networks-based algorithm, logistic regression, CART, C4.5, J48, (BayesNet), SimpleLogistic, JRip, RandomForest, Logistic regression analysis, ICRM2 [6]. However, under the classification techniques, Decision Tree is highly used by researchers due to its simplicity and comprehensibility to uncover small or large data structure and predict the value [2].

Other techniques such as survival analysis, which provides various mechanisms to handle such censored data problems that arise in modeling such longitudinal data which occurs ubiquitously in various real-world application domains were presented [13]. Ameri et al, developed a survival analysis framework for early prediction using the Cox proportional hazards model (Cox) and applied time-dependent Cox (TD-Cox), which captures the time-varying factors and can leverage this information to provide more accurate prediction of student dropout using the dataset of students enrolled at Wayne State University (WSU) starting from 2002 until 2009 [7]. Besides, other researchers proposed a new data transformation model, which is built upon the summarized data matrix of link-based cluster ensembles (LCE) using educational dataset obtained from the operational database system at Mae Fah Luang University, Chiang Rai, Thailand. Like several existing dimension reduction techniques such as principal component analysis (PCA) and kernel principal component analysis (KPCA), this method aims to achieve high classification accuracy by transforming the original data to a new form. However, the common limitation of these new techniques is the demanding time complexity, such that it may not scale up well to a very large dataset. Whilst worst-case traversal time (WCT-T) is not quite for a highly time-critical application, it can be an attractive candidate for those qualities-led works, such as the identification of those students at risk of underachievement [5]. Furthermore, matrix factorization as a clustering machine learning method that can accommodate framework with some variations was presented [10]. In this study, two classes of methods for building the prediction models were described. 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. The mentioned approach was applied to the dataset generated from George Mason University (GMU) transcript data, University of Minnesota (UMN) transcript data, UMN LMS data, and Stanford University MOOC data [11]. However, 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.

In this paper, we present a thorough analysis of four commonly used machine learning algorithms on Uwezo data on learning (http://www. twaweza.org/go/uwezo-datasets) in Tanzania with the aim to provide data-driven algorithm recommendations to current researchers on the topic. This is publicly available national wide dataset in Tanzania which was generated from developing country and therefore reflects the local context. Using new sources of student level dataset from Tanzania as a case study, we employ a comprehensive validation and enhancement to existing algorithms and apply additional machine learning approaches to improve their predictive power. Specifically, we take a detailed analysis of selected popular algorithms and analyse their performance on the dataset by first applying data pre-processing and feature engineering techniques which are very critical states for building high performance dropout prediction algorithm. This was followed by a rigorous comparison of selected machine learning algorithms using evaluation methods as proposed by [18] in which the best performing algorithms were selected. Further, we empirically quantified the effect of hyper parameters (i.e. algorithm parameters) tuning and ensemble techniques for the selected algorithms with an aim to further improve their performance.

In summary, the main objective of this study focused on applying machine learning techniques for predicting student dropout. In order to attain this objective, the following three tasks were performed respectively:

- Building the model and analyzed the performance.
- Tuning models that performed well and employed ensemble approach to improve the predictive performance.

- Evaluated model performance using three metrics: geometric mean (Gm), F-measure (Fm) and adjusted geometric mean (AGm).

2.Materials and methods

2.1Dataset descriptions and pre-processing

Data pre-processing includes data cleaning, normalization, transformation, feature extraction and selection, etc., and the product of data pre-processing is the final training set. In selection, relevant target data are selected from retained data (typically very noisy) and subsequently pre-processed. This goes hand in hand with the integration from multiple sources, filtering irrelevant content and structuring of data according to a target tool [19]. In developing a generalized algorithm, data pre-processing 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 a considerable amount of processing time in ML problems.

In this paper, Uwezo data on learning at the country level in Tanzania which was collected in 2015 was used. This dataset was collected by Twaweza organization with the aim of assessing children's learning levels across hundreds of thousands of households in East Africa. The dataset was cleaned by removing information from the data that could lead to individuals or specific villages being located by end-users. Village id column was removed, since it was not required in experimental stage. Various approaches have been identified in handling missing values, outliers, data and numeric values [20]. In this study, we converted data samples to numerical values and performed PCA for handling outliers. Missing values were replaced using medians and zeros.

In this dataset, we identified the following columns have missed values as described in *Table 1*: Pupil Teacher Ratio (PTR), Pupil Classroom Ratio (PCR), Girl's Pupil Latrines Ratio (GPLR), Boy's Pupil Latrines Ratio (BPLR), Parent Teacher Meeting Ratio (PTMR), Main source of household income (Income), Enumeration Area type (EAarea), Parent who check his/her child's exercise book once in a week (PCCB), Parent who discuss his/her child's progress with teacher last term (PTD), Student who did read any book with his/her parent in last week (SPB), School has girl's privacy room (SGR), Household meals per day (MLPD). On handling missing values, PTR, PCR, GPLR, BPLR were imputed with medians and PTMR, Income, EAarea, PCCB, PTD, SPB, SGR, MLPD were imputed with zeros. We encoded the nominal features to conform with Scikit-learn and change the dropout code: 1 to represent not drop and 0 to represent dropout.

The dataset consists of 18 features as described in *Table 2* and approximately 61340 samples. Since our target variable is dropout, we checked the distribution of this variable in the dataset and observed that there was an imbalance for target variable with only 1.6% dropout as shown in *Figure 1*. Data imbalance is a serious problem which can be considered during pre-processing stage [21]. This happens when one class is under-represented relative to another [22, 23]. Classification of imbalance dataset is common in the field of student retention, mainly because the number of registered students is always larger than the dropout students. Several re-sampling techniques such as under-sampling, over-sampling and hybrid methods can be applied to handle this problem [24].

Table 1 Features with missing values

No.	Feature description	Type of data
1.	Main source of household income (Income)	Multinominal
2.	Boy's pupil latrines ratio (BPLR)	Numerical
3.	School has girl's privacy room (SGR)	Binominal
4.	Parent who check his/her child's exercise book once in a week (PCCB)	Binominal
5.	Household meals per day (MLPD)	Multinominal
6.	Student who did read any book with his/her parent in last week (SPB)	Binominal
7.	Parent who discuss his/her child's progress with teacher last term (PTD)	Binominal
8.	Enumeration Area type (EAarea)	Multinominal
9.	Girl's pupil latrines ratio (GPLR)	Numerical
10.	Parent teacher meeting ratio (PTMR)	Numerical
11.	Pupil classroom ratio (PCR)	Numerical
12.	Pupil teacher ratio (PTR)	Numerical

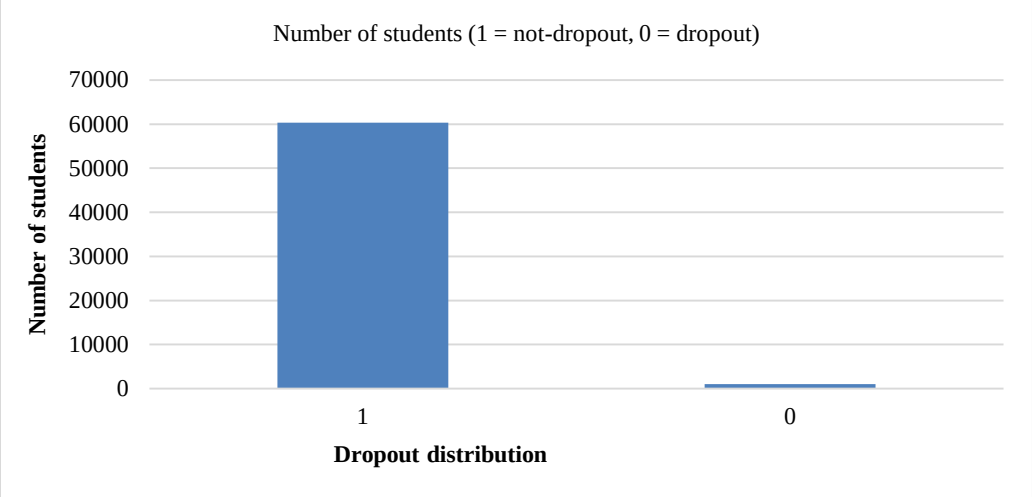

Figure 1 Dropout distribution of the dataset

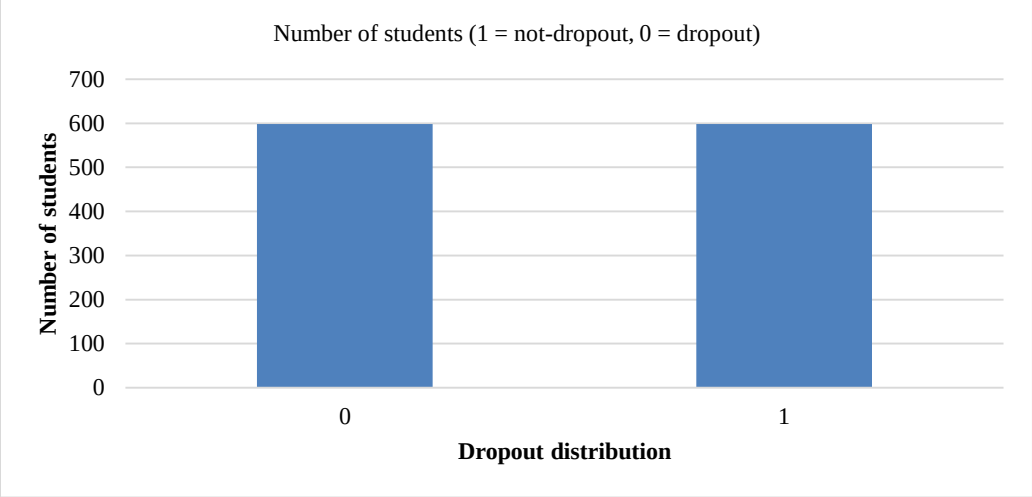

Figure 2 Dropout distributions (under-sampling)

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 [25, 26]. Over-sampling method create a superset of the original data set 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 [27−29]. Hybrids method such as synthetic minority oversampling (SMOTE).

Technique) combines both under-sampling and over-sampling approaches [30, 31]. In this case, no sampling, under-sampling and over sampling (SMOTE-ENN) and random under sampling technique as implemented in Imbalanced-Learn were applied as demonstrated in *Figure 2* and *3*. SMOTE-ENN combines over and under sampling using SMOTE and edited nearest neighbour (EN) to generate more minority class in order to reinforce its signal [32] and random under sampler is a fast and easy way to balance the minority class by randomly selecting a subset of data for the targeted classes [33].

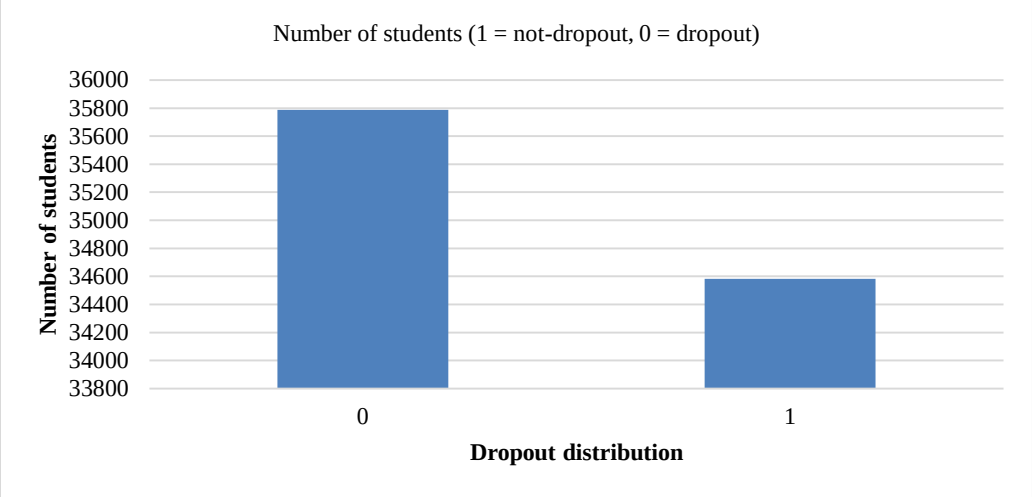

Figure 3 Dropout distributions (SMOTE-ENN)

Table 2 Summary of all features

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

2.2Feature selection

Feature selection is one of the useful approaches in data pre-processing for finding accurate data models [34]. The experiments aim at identifying the contribution of each feature on the prediction performance by automatically selecting features that are most relevant to the dropout predictive modeling. This was accomplished by measuring permutation of the feature importance score (pfi) as defined in Equation 1.

$$pfi = P_b - P_s \quad (1)$$

Where:

- P_b is the base performance metric score
- P_s is the performance metric score after shuffling

The score was intended to measure the impact of each feature on the model performance by permuting the values of each feature and measuring how much the permutation decreases the model performance. In this experiment, important measures were randomly computed by permutations of feature value depending on the contribution of each feature to the predictive performance of a model. This was followed by measuring the deviation after permuting the values of that feature using Gm as an evaluation metric. The overall experiment was conducted by creating a random permutation of feature through the shuffling of values and evaluates model performance, which was done iteratively for each feature, one at a time to observe the list of the feature variables and their corresponding importance scores. The

importance score is defined as the reduction in performance after shuffling the feature values. When evaluation metric was used to measure the accuracy of the prediction, a higher value implies the feature is more important.

The results presented in *Figure 4*, show that student gender (Gender), PCCB, MLPD, SPB, PTD and Student age (Age) have a strong contribution to the dropout prediction performance. Thereafter, the same experiment was repeated using six well performed features obtained in the previous experiment. The results show clearly that a student's gender has a strong contribution to the dropout prediction performance as presented in *Figure 5*. These experimental results, support researchers' findings on dropout rate with gender association [35].

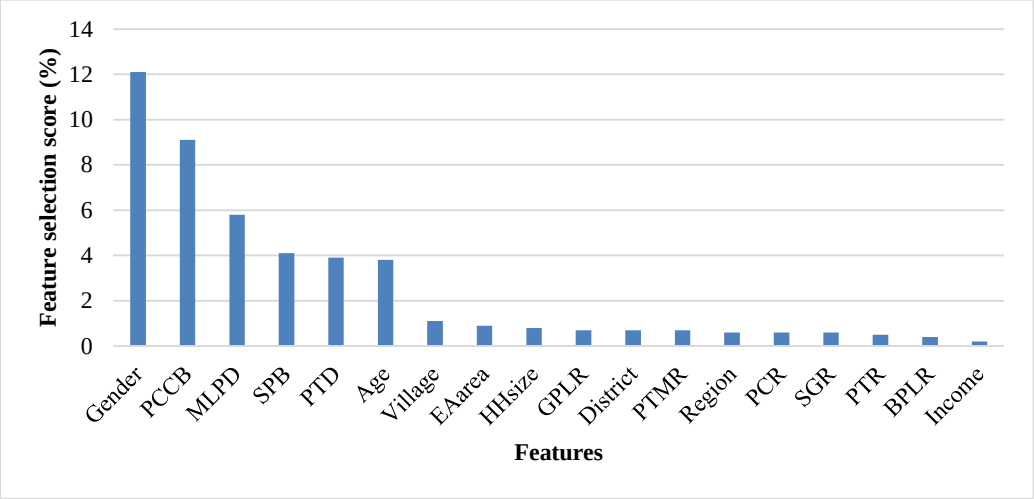

Figure 4 Feature selection with all features

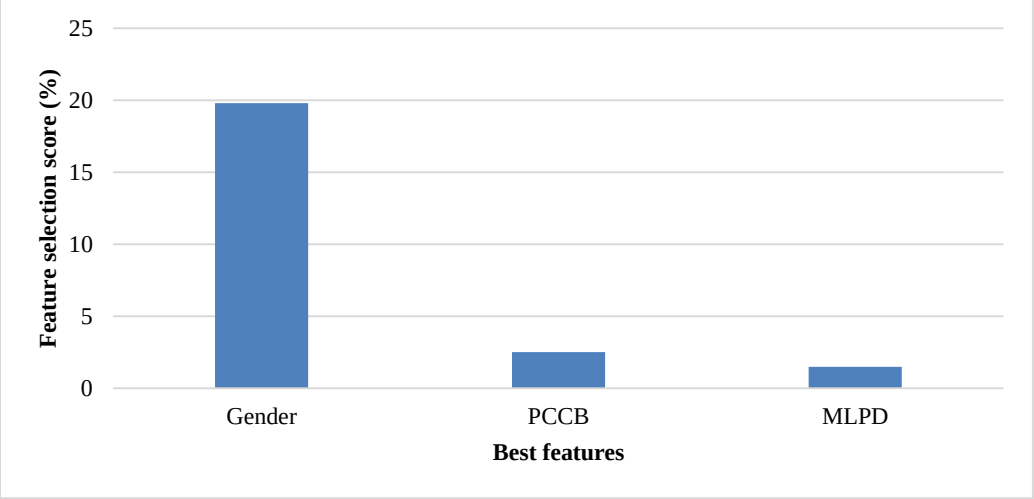

Figure 5 Feature selection with best features

2.3Experimental procedures

In this study, the dataset was separated into a train (60%), validation (20%) and test (20%). We applied the sampling techniques to the training set and conducted the first experiment of building the model. The model was built using 60% of training set and then 20% of validation set was used to validate the model performance. This was followed by the second experiment of tuning the well performed models and employed ensemble techniques in order to improve the predictive performance. We then combined train and validation sets formulate a big train set and applied sampling technique to this train set. Thereafter, we evaluate the model using unseen 20%

of test set in order to observe how the model will behave in a real environment which is an imbalance. The overall experimental procedure is summarized in *Figure 6* wherein each experiment stratified k-fold cross validation was used. In this experiment, k=5 fold out-of-bag overall cross validation was used and the entire process involves executing all selected classification algorithms in which all executions were repeated 5 times using different train/test/validation 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.

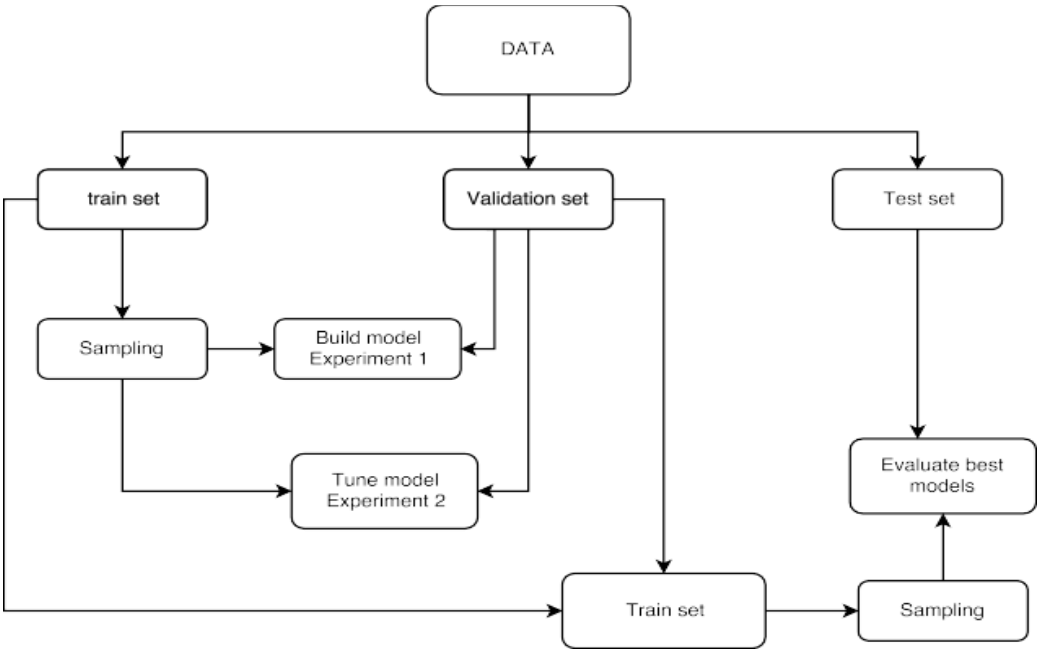

Figure 6 Experiment procedure

2.4Evaluation metrics

In measuring the performance of student dropout algorithms, several researchers use various evaluation metrics [1, 7, 8]. With respect to evaluation measures, we used Gm, Fm and AGm as evaluation criteria. The Gm is a measure of the ability of a classifier to balance TPrate (sensitivity) and TNrate (specificity) [36] as defined in Equation 2. This measure is maximum when the true positive rate (TPrate) and the true negative rate (TNrate) are equal. Furthermore, in order to ensure TPrate to the changes in the positive predictive value (precision) than in TPrate, Fm was used as defined in Equation 3. This is the weighted harmonic mean of the TPrate and precision [7, 37, 38]. Besides, AGm as defined in Equation 4 was used to obtain the highest TPrate without decreasing too much the TNrate [18].

$$G_m = \sqrt{TPrate \,.\, TNrate} \quad (2)$$

$$F_m = 2PPV.\frac{TPrate}{PPv+Tprate} \quad (3)$$

$$AG_m = \begin{cases} \frac{G_m+TNrate\,.\,(FP+TN)}{1+FP+TN} & if\ Tprate > 0, \\ 0 & if\ Tprate = 0 \end{cases} \quad (4)$$

Where:

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

3.Results

3.1Experiment 1: model selection

The aim of this experiment was to identify classifier with the best performance for this problem. In this phase, selection of classifiers was based with all domains, including linear, ensemble, instance and neural network classifiers with consideration of the classification and nature of the

dataset. Linear models were represented by a logistic regression classifier (LR), ensemble models were represented by random forest (RF), instance model was represented by K-Nearest-Neighbors (KNN) and neural networks models were represented by a multilayer perceptron (MLP). The experiment was repeated for three different cases: SMOTE-ENN and results in both three cases are presented. Results were presented on a separate graph based on the scale of evaluation metrics. For GM and Fm metrics with scale range between 0 and 1, the results were combined in the same graphs (*Figure 7-9*), while AGm metric with scale range above 0 were presented in a separate graph (*Figure 10-12*). To select the best classifiers, validation results were considered because it gives an estimate on how the classifier will perform on actual dataset which is an imbalance. From the result presented in *Figure 7 and 10*, two classifiers: LR and MLP show better generalization results. They show better validation results for the three metrics used. Considering the case when under-sampling is used as observed in *Figure 8 and 11*, all classifiers have considerably the same generalization results for both metrics. The experiment conducted without sampling as presented in *Figure 9 and 12* reveal that, only LR classifier show better performance than others. However, the score rates are less than 1 for AGm as compared to when LR is used with oversampling case. Therefore, for the next experiment the following two classifiers: LR and MLP were considered with oversampling case.

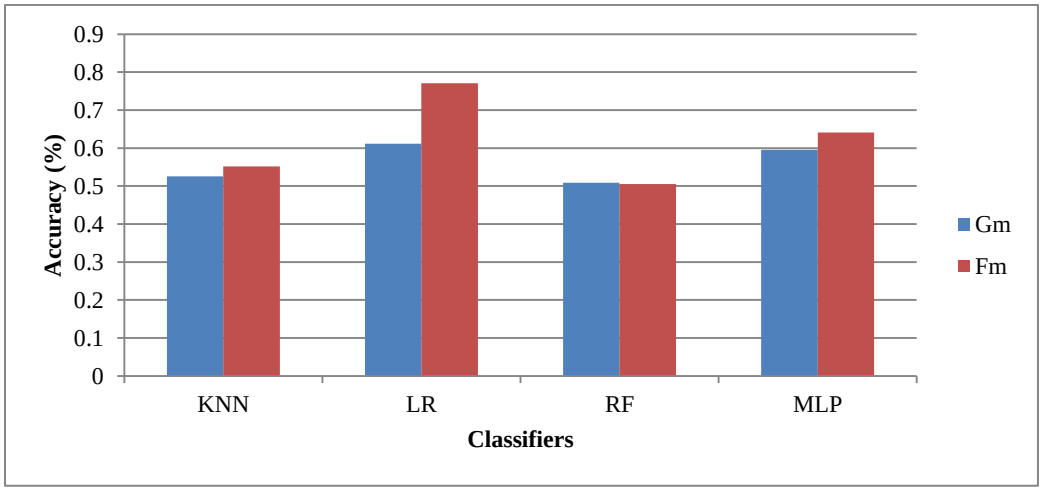

Figure 7 Validation results (over-sampling)

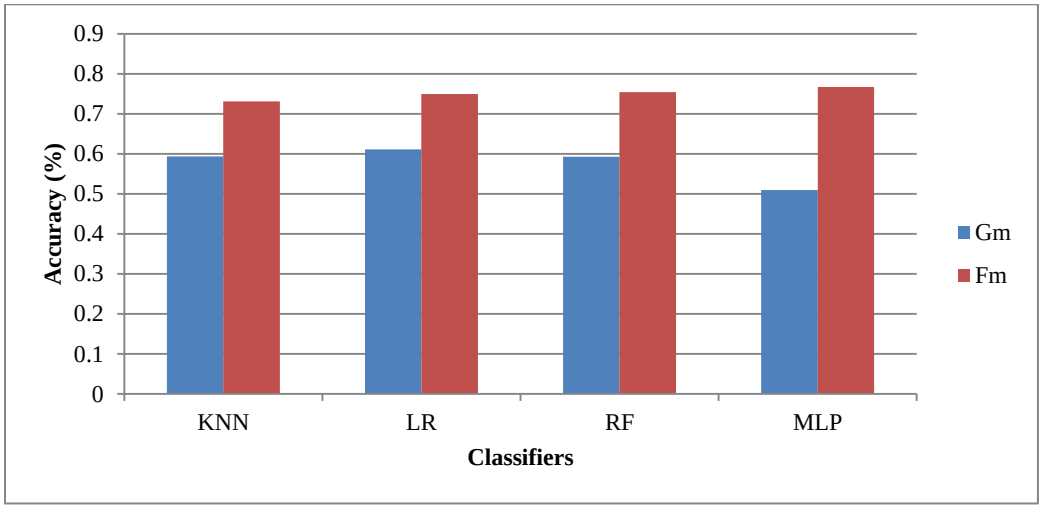

Figure 8 Validation results (under-sampling)

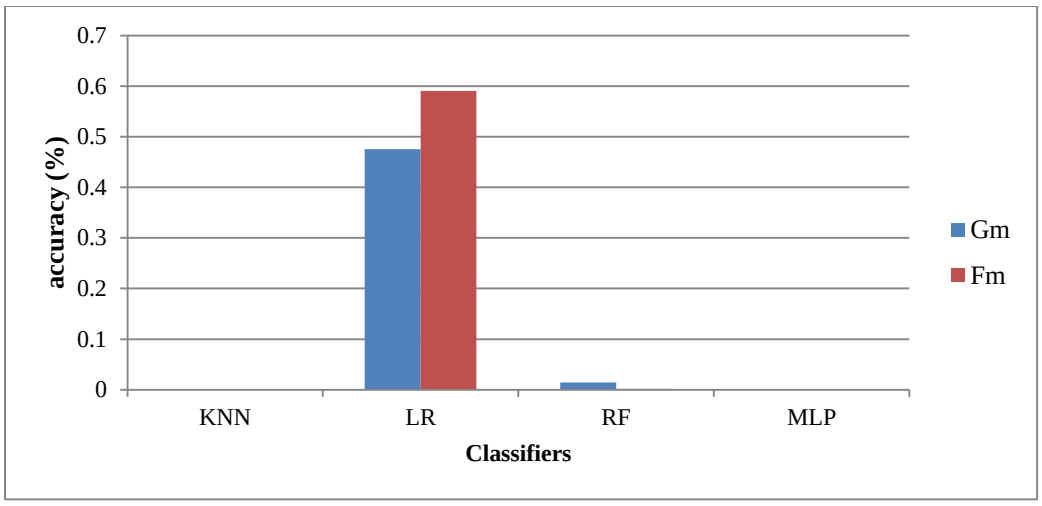

Figure 9 Validation results (no sampling)

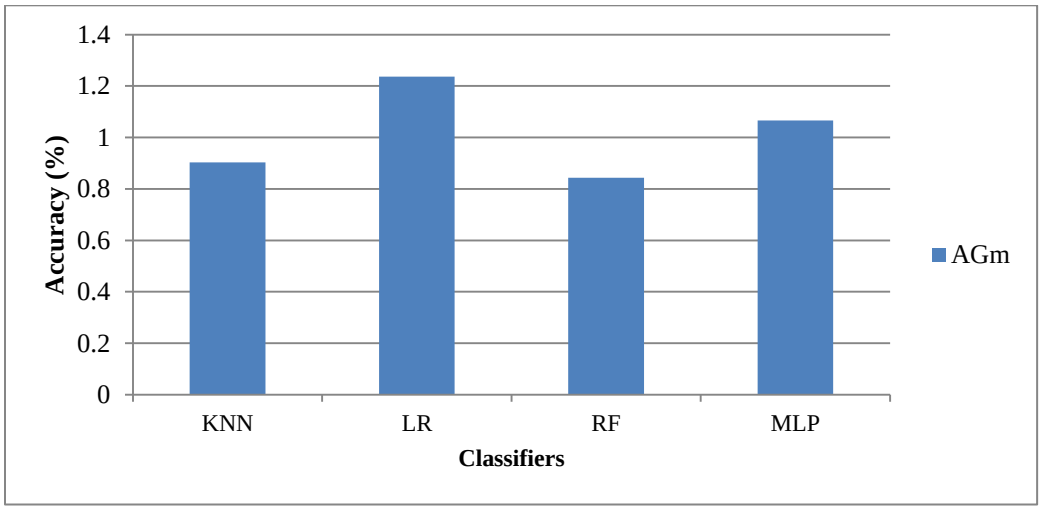

Figure 10 Validation results (over-sampling)

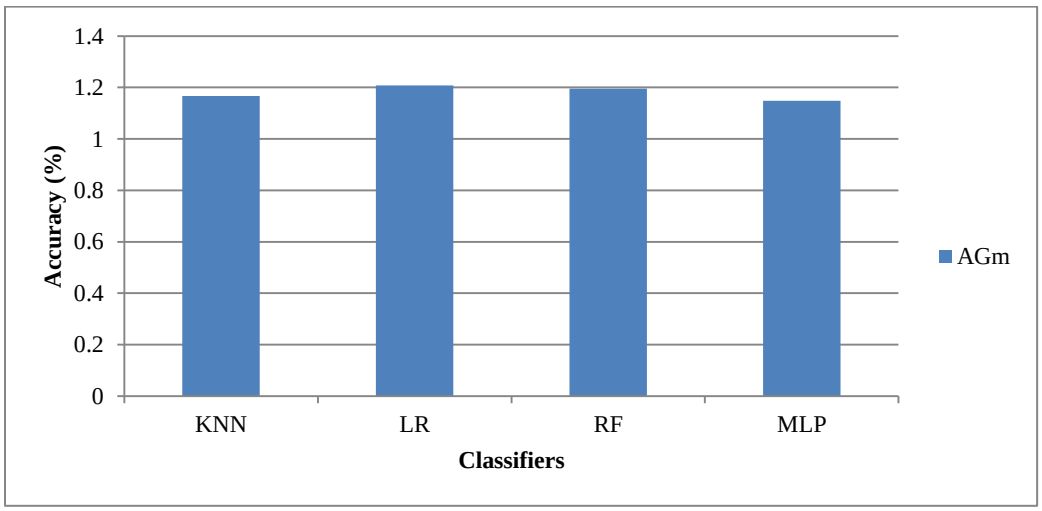

Figure 11 Validation results (under-sampling)

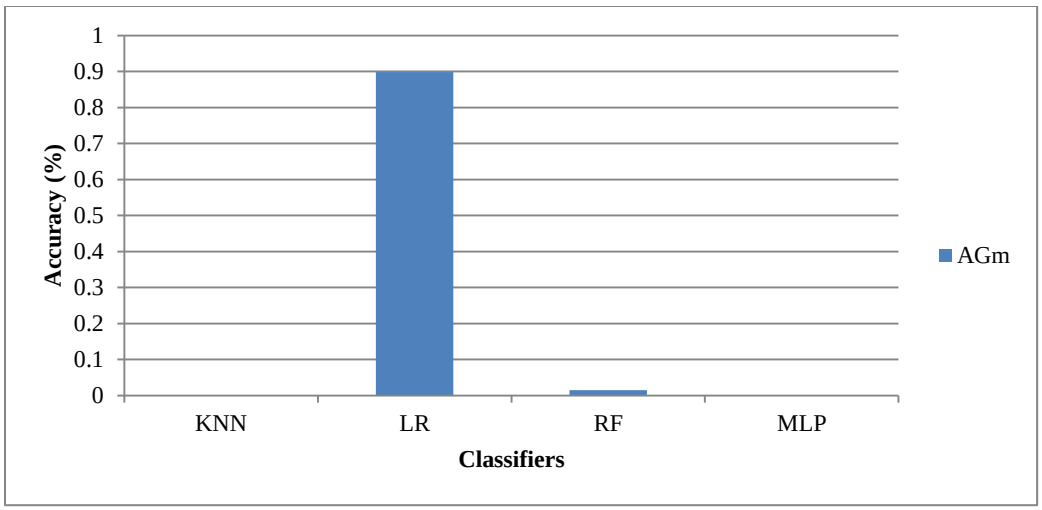

Figure 12 Validation results (no sampling)

3.2Experiment 2: Hyper-parameter optimization

This experiment aims to show the significance of hyper-parameter tuning on improving predictive performance. It involved a combination of hyper-parameter values for a machine learning algorithm that performs the best as measured on a validation dataset. Most ML algorithms contain several hyper parameters that can affect performance significantly (for example, the number hidden layers in MLP classifier) [39, 40]. In this experiment two selected classifiers were tuned: LR and MLP to further improve their performance. The grid search approach was employed so as to set a grid of hyper parameter values and for each combination, train a model and score on the validation data. In order to evaluate each combination of hyper parameter values, we scored them on a validation set. Hyper-parameter tuning via cross-validation was implemented using 5-fold cross validation and identified the best parameters for each classifier as presented in *Table 3*.

The experimental results allow us to measure the extent to which hyper parameter tuning improves each algorithm's performance compared to its baseline settings. Ensemble technique was employed in order to improve the overall predictive performance of the models. This method is one of the popular approaches for improving machine learning algorithms.

Table 3 Model parameters

Classifier	Parameter
LR	fit intercept: True, tol:1, C:0.001, Penalty:'ll'
MLP	solver:'adam', learning rate int:0.001, shuffle: True, hidden layer size:10, alpha:1, early stopping: True

The ensemble approach creates multiple models and then combines them to produce improved results. Several ensemble techniques such as bagging, boosting and voting have been extensively used in the literature [41, 42]. For this problem, the voting ensemble technique was appropriate. Voting (stacking) was employed by soft combined the two tuned classifiers LR2 and MLP2. The tuned classifiers where then trained on the new training set obtained by combining validation and training set used in previous experiments. On evaluating the generalization performance, models were tested on unseen tested data and evaluation of the models was done by comparing validation results.

Table 4 Experiment 2: results

		LR	LR2	MLP	MLP2	ENB
Validation scores	Gm	0.724	0.726	0.613	0.711	0.735
	AGm	1.261	1.372	1.211	1.324	1.370
	Fm	0.841	0.894	0.723	0.827	0.891
Test scores	Gm	0.721	0.783	0.621	0.706	0.779
	AGm	1.320	1.332	1.278	1.281	1.335
	Fm	0.823	0.831	0.726	0.732	0.847

Results presented in *Table 4* reveal that, performance of the tuned algorithms (LR2 and MLP2) was improved compared to untuned algorithms (LR and MLP). Furthermore, the stacking classifier (ENB) shows considerably better validation and test results followed by the tuned logistic regression model (LR2).

4.Discussion

Although a number of literatures have shown the feasibility of explaining student dropout, few works have actually attempted to predict student dropout. In this study, we use machine learning techniques that are able to automatically identify features that are relevant. With the right model, it was possible to predict students' dropout as well as explain the variables that are likely to be useful in the prediction. We achieved this by employing the ensemble classifier that tends to do better than a single individual classifier. This classifier which was produced by soft combining the tuned LR2 and MLP2 achieved better results followed by the tuned LR2. The machine learning approach of combining multiple classifiers has been proposed for improving predictive performance [43], and generates better results [44]. Furthermore, we observed student gender as the leading feature which shows high contribution to the student dropout problem and hyper parameter tuning improves algorithm performance. Compared to the results presented by [2] as described in *Table 5*, J48 showed better results on proposing student advising model for enhancing students' academic performance and decreasing dropout. Three decision tree classification algorithms, namely J48, random tree and reduces error pruning (REP) tree were used in a real dataset representing students' records in a managerial higher institute in Giza Egypt. The approach used in our presented study, focused on analyzing four supervised learning classifiers that represent linear, ensemble, instance and neural networks rather than focusing only on decision tree classification algorithms.

Furthermore, on investigating prediction algorithm for academic performance on tackling the problem of student dropout [1]. LR achieved the highest performance on comparison results of classification performance for all the six classifiers which are LR, MLP, sequential minimal optimization (SMO), naive Bayes (NB), J48 and RF using six metrics on the dataset collected from rural and peri-urban primary schools in Kenya as shown in *Table 6*. LR achieved better results in our presented study.

Table 5 Classification results for three algorithms [Comparison from [2]]

Algorithm	**Time (sec)**	**Model evaluation**			
		Correctly classified		**Incorrectly classified**	
		#	**%**	**#**	**%**
J48	0.05	7081	87.64	999	12.36
Random tree	0.02	7065	87.43	1015	12.56
REP tree	0.03	7065	87. 44	1015	12.56

Table 6 A comparison of the classifiers' performance using the six selected metrics [Comparison from [1]]

Model	**Recall**	**Specificity**	**ROC**	**F-Measure**	**Kappa**	**RMSE**
LR	0.924	0.686	0.887	0.897	0.6345	0.3375
MLP	0.873	0.660	0.851	0.865	0.5407	0.4124
SMO	0.911	0.703	0.807	0.894	0.6309	0.3893
NB	0.701	0.801	0.846	0.784	0.4403	0.4264
J48	0.905	0.670	0.822	0.884	0.5941	0.3720
RF	0.907	0.684	0.870	0.888	0.6082	0.3471

5.Conclusions

In this paper, a case study has been presented that shows application of machine learning approach on addressing the problem of student dropout. Four supervised classification algorithms were empirically assessed 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 the data imbalanced problem. The two classifiers LR and MLP have proven superior to all the other classifiers by achieving highest performance metrics when over-sampling technique was employed. Furthermore, the results show that the hyper-parameter tuning improves each algorithm's

performance compared to its baseline settings and stacking these classifiers improves the overall predictive performance. Also, the contribution of each feature on the prediction performance with student gender being the leading feature was shown. For future work, we plan to explore different datasets so as to show and compare results of different train, test and validation and evaluate several imbalance techniques for student dropout prediction using more measures for results comparison. This will include extending the experiment by applying under sampling approach with penalized models on resolving the imbalance issue. Besides, we will generalize the study and add more features so as to evaluate feature subsets for better understanding of the underlying process.

Acknowledgment

The authors would like to thank the African Development Bank (AfDB), Data for Local Impact (DLi), Eagle Analytics Company, Late Dr. Yaw-Nkansah Gyekye and Anthony Faustine for supporting this study.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Mgala M. Investigating prediction modelling of academic performance for students in rural schools in Kenya (Doctoral dissertation, University of Cape Town). 2016.

[2] Mohamed MH, Waguih HM. A proposed academic advisor model based on data mining classification techniques. International Journal of Advanced Computer Research. 2018; 8(36):129-36.

[3] KH, Van Der Schaar M. A machine learning approach for tracking and predicting student performance in degree programs. IEEE Journal of Selected Topics in Signal Processing. 2017; 11(5):742-53.

[4] Feng W, Tang J, Liu TX. Understanding dropouts in MOOCs. Association for the Advancement of Artificial Intelligence. 2019.

[5] Iam-On N, Boongoen T. Generating descriptive model for student dropout: a review of clustering approach. Human-Centric Computing and Information Sciences. 2017; 7(1).

[6] Kumar M, Singh AJ, Handa D. Literature survey on educational dropout prediction. International Journal of Education and Management Engineering. 2017; 7(2):8-19.

[7] Ameri S, Fard MJ, Chinnam RB, Reddy CK. Survival analysis based framework for early prediction of student dropouts. In proceedings of the ACM international on conference on information and knowledge management 2016 (pp. 903-12). ACM.

[8] Aulck L, Velagapudi N, Blumenstock J, West J. Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364. 2016.

[9] Chen Y, Chen Q, Zhao M, Boyer S, Veeramachaneni K, Qu H. DropoutSeer: visualizing learning patterns in massive open online courses for dropout reasoning and prediction. In conference on visual analytics science and technology 2016 (pp. 111-20). IEEE.

[10] Hu Q, Polyzou A, Karypis G, Rangwala H. Enriching course-specific regression models with content features for grade prediction. In international conference on data science and advanced analytics 2017 (pp. 504-13). IEEE.

[11] Elbadrawy A, Polyzou A, Ren Z, Sweeney M, Karypis G, Rangwala H. Predicting student performance using personalized analytics. Computer. 2016; 49(4):61-9.

[12] Iqbal Z, Qadir J, Mian AN, Kamiran F. Machine learning based student grade prediction: a case study. arXiv preprint arXiv:1708.08744. 2017.

[13] Wang W, Yu H, Miao C. Deep model for dropout prediction in MOOCS. In proceedings of the international conference on crowd science and engineering 2017 (pp. 26-32). ACM.

[14] Hamedi A and Dirin A. A Bayesian approach in students' performance analysis. International conference on education and new learning technologies. 2018.

[15] https://icsh.es/2017/11/12/i-congreso-internacional-multidisciplinario-de-educacion-superior/. Accessed 26 October 2018.

[16] Hung JL, Wang MC, Wang S, Abdelrasoul M, Li Y, He W. Identifying at-risk students for early interventions-a time-series clustering approach. IEEE Transactions on Emerging Topics in Computing. 2017; 5(1):45-55.

[17] Młynarska E, Greene D, Cunningham P. Time series clustering of MOODLE activity data. In Irish conference on artificial intelligence and cognitive science University College Dublin, Dublin, Ireland, 2016.

[18] Yan J, Han S. Classifying imbalanced data sets by a novel re-sample and cost-sensitive stacked generalization method. Mathematical Problems in Engineering.2018.

[19] Alasadi SA, Bhaya WS. Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences. 2017; 12(16):4102-7.

[20] Shahul S, Suneel S, Rahaman MA, Swathi JN. A study of data pre-processing techniques for machine learning algorithm to predict software effort estimation. Imperial Journal of Interdisciplinary Research. 2016; 2(6):546-50.

[21] Krawczyk B. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence. 2016; 5(4):221-32.

[22] Galar M, Fernández A, Barrenechea E, Bustince H, Herrera F. New ordering-based pruning metrics for ensembles of classifiers in imbalanced datasets. In proceedings of the international conference on

computer recognition systems 2016 (pp. 3-15). Springer, Cham.
[23] Borowska K, Topczewska M. New data level approach for imbalanced data classification improvement. In proceedings of the international conference on computer recognition systems 2015 (pp. 283-94). Springer, Cham.
[24] Rout N, Mishra D, Mallick MK. Handling imbalanced data: a survey. In international proceedings on advances in soft computing, intelligent systems and applications 2018 (pp. 431-43). Springer, Singapore.
[25] Saini AK, Nayak AK, Vyas RK. ICT Based Innovations. Proceedings of CSI. 2015.
[26] Dattagupta SJ. A performance comparison of oversampling methods for data generation in imbalanced learning tasks (Doctoral dissertation). 2017.
[27] Stefanowski J. On properties of undersampling bagging and its extensions for imbalanced data. In proceedings of the international conference on computer recognition systems 2016 (pp. 407-417). Springer, Cham.
[28] Moreno MF. Comparing the performance of oversampling techniques for imbalanced learning in insurance fraud detection (Doctoral dissertation). 2017.
[29] Santoso B, Wijayanto H, Notodiputro KA, Sartono B. Synthetic over sampling methods for handling class imbalanced problems: a review. In IOP conference series: earth and environmental science 2017 (p. 012031). IOP Publishing.
[30] Skryjomski P, Krawczyk B. Influence of minority class instance types on SMOTE imbalanced data oversampling. In first international workshop on learning with imbalanced domains: theory and applications 2017 (pp. 7-21).
[31] Ahmed S, Mahbub A, Rayhan F, Jani R, Shatabda S, Farid DM. Hybrid methods for class imbalance learning employing bagging with sampling techniques. In international conference on computational systems and information technology for sustainable solution 2017 (pp. 1-5). IEEE.
[32] Douzas G, Bacao F. Geometric SMOTE: effective oversampling for imbalanced learning through a geometric extension of SMOTE. arXiv preprint arXiv:1709.07377. 2017.
[33] Elhassan T, Aljurf M. Classification of imbalance data using tomek link (T-Link) combined with random under-sampling (RUS) as a data reduction method. Global Journal of Technology and Optimization. 2016, S1: 111.
[34] Khaldy MA, Kambhampati C. Resampling imbalanced class and the effectiveness of feature selection methods for heart failure dataset. International Robotics & Automation Journal. 2018; 4(1):1-10.
[35] Kim D, Kim S. Sustainable education: analyzing the determinants of university student dropout by nonlinear panel data models. Sustainability. 2018; 10(4):1-18.
[36] Márquez-Vera C, Cano A, Romero C, Noaman AY, Mousa Fardoun H, Ventura S. Early dropout prediction using data mining: a case study with high school students. Expert Systems. 2016; 33(1):107-24.
[37] Rovira S, Puertas E, Igual L. Data-driven system to predict academic grades and dropout. PLoS one. 2017; 12(2):e0171207.
[38] Aulck L, Aras R, Li L, L'Heureux C, Lu P, West J. STEM-ming the tide: predicting STEM attrition using student transcript data. arXiv preprint arXiv:1708.09344. 2017.
[39] Rojas-Domínguez A, Padierna LC, Valadez JM, Puga-Soberanes HJ, Fraire HJ. Optimal hyper-parameter tuning of SVM classifiers with application to medical diagnosis. IEEE Access. 2018; 6:7164-76.
[40] Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 1804:e1301.
[41] Dalvi PT, Vernekar N. Anemia detection using ensemble learning techniques and statistical models. In international conference on recent trends in electronics, information & communication technology 2016 (pp. 1747-51). IEEE.
[42] Feng W, Huang W, Ren J. Class imbalance ensemble learning based on the margin theory. Applied Sciences. 2018; 8(5):815.
[43] Abuassba AO, Zhang D, Luo X, Shaheryar A, Ali H. Improving classification performance through an advanced ensemble based heterogeneous extreme learning machines. Computational Intelligence and Neuroscience. 2017.
[44] Afolabi LT, Saeed F, Hashim H, Petinrin OO. Ensemble learning method for the prediction of new bioactive molecules. PloS one. 2018; 13(1):e0189538.

Neema Mduma is a PhD fellow in the department of Information and Communication Science and Engineering (ICSE) at the Nelson Mandela African Institution of Science and Technology (NM-AIST). Her focus is on supporting education and currently she is conducting a study on developing a machine learning approach for predicting student dropout.
Email: mduman@nm-aist.ac.tz

Khamisi Kalegele is a Lecturer and Researcher at the Tanzania Commission of Science and Technology (COSTECH). He graduated with a PhD in Information Sciences from Tohoku University, Japan in 2013; MEng in Computer Science from Ehime University in Japan and BSc Computer Engineering and IT from University of Dar Es Salaam. His research areas are Data Science, E-health and Machine Learning in Education.
Email: kalegs03@gmail.com

Dina Machuve is a Lecturer and Researcher at the Nelson Mandela African Institution of Science and Technology (NM-AIST) in Tanzania. She graduated with a PhD in Information and Communication Science and Engineering from NM-AIST in 2016, and with a MS in Electrical Engineering from Tennessee Technological University, USA in 2008 and BSc Electrical Engineering degree from the University of Dar Es Salaam in 2001. She serves on the organizing committee of Data Science Africa, an organization that runs an annual data science and machine learning summer school and workshop in Africa. Her research interests are Data Science, Bioinformatics, Agriculture Informatics on Food Value Chains and STEM Education in Schools.
Email: dina.machuve@nm-aist.ac.tz