INTERNATIONAL JOURNAL OF ADVANCED STUDIES IN COMPUTER SCIENCE AND ENGINEERING
IJASCSE VOLUME 8 ISSUE 02, 2019
02/28/2019

A Machine Learning Approach to Dropout Early Warning System Modeling

[1]Isiaka R.M., [2]BabatundeR.S., [3]AjaoF.J. and [4]Abdulsalam S.O.
[1,2,3,4]Department of Computer Science, Kwara State University, Malete. Nigeria.

ABSTRACT
Dropout has been identified as a social problem with an immediate and long term consequences on the socio-economic development of the society. A major element of a conceptual framework for dropout monitoring and control system is the prevention component. The component specified the need for automatic dropout early warning system. This paper presents the procedure for building the adaptive model, that is the core element for the realization of the prevention component of the framework. The model development is guided by the knowledge of the domain experts. An experimental approach was used to identify the K-Nearest Neighbors (KNN) as the best of the six algorithms considered for the adaptive models. Other algorithms explored are Logistic Regression (LR), Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), Gaussian Naive Bayes (NB) and Support Vector Machines (SVM). Historical dataset from a State University in the North Central region of Nigeria was used for building and testing the adaptive model. The average performance of the model on full attributes and the first year academic attributes compared favourably with 0.98 Precision, Recall and F1-Scores. Conversely, the Precision, Recall and F1-Scores on the entry attributes are 0.90, 0.95 and 0.92 respectively. The implication of this is that the academic performance is a very significant factorfor dropping out of the educational system. The developed model can provide an early signal on the students with the propensity to dropout, thereby serving as an advisory system for the students, parents and school managementtowards curtailing the menace.

Keywords: dropout, adaptive-model, data mining, machine learning prediction

I. INTRODUCTION

Issues related to the determination of students dropout causative factors, possible prevention, and control strategies have become attractive to researchers in Educational Data mining. The multifaceted and multi-dimensional consequences of dropouts are very conspicuous in the literature. It affects the income and reputations of Educational Institutions. It constitutes loss of investments and dampens parents and sponsors' confidence in the system. Furthermore, it derails students' dreams and degrades societal expectations. Greater percentages of dropouts occur at the early stage of the educational pursue. It was asserted in[8], that, more than fifty percent (50%) of student attrition occur among freshers. It is therefore essential to galvanise strategies towards early detection of students with the propensity of losing studentship before graduation. Consequently, in [9] a framework for dropout monitoring and management system was proposed. It was hinged on the belief that the successes of Higher Education Institution in managing student's retention determine their funding, ranking and Alumni contributions.

The framework is a conceptual framework for dropout monitoring and control in Nigerian Universities. It was built on theories of Attribution, Social- Capital, and Engagement. The three components of the framework are causes, prevention and control. The prevention component, which is the second component, specified the need for automatic dropout early warning system. Consequently, this paper reveals the procedure for building the adaptive model which is a core element for the realization of the prevention component of the framework. A basic requirement for creating an adaptive model is an historical dataset. Hence, a dataset for a state university in the North Central region of Nigeria was used for building and testing the adaptive model.

The rest of this paper is structured as follows; The Section II presents previous researches related to Educational Data mining. The methodology is discussed in Section III, it comprises of the approaches

used for obtaining the dataset, the data exploration and the data visualization. The section also present the procedures employed in the actual modeling and evaluation of the system. The conclusion and future goals are presented in Section IV.

II. RELATED WORKS

Educational Data Mining is a fast-growing interdisciplinary research field [3]. It has been applied to evaluation of student performance, course recommendation and identification of learning pattern for implementation of personalized learning [4]. It has also proved to be useful for discovery of at-risk student [8].

It was stressed in [10] that the classical analytical approach is no more sufficient in modern information systems; instead, business intelligence tools are required. Therefore, there was the need to comprehend data using statistical data processing technique. This was demonstrated in the work and followed by model building according to SEMMA methodology using Logistic Regression, Decision Tree and Neural Networks. SEMMA signifies Sampling, Exploring, Modifying, Modeling and Assessment.

Several other efforts have been made applying Machine Learning and Data Mining techniques to studying academic performance and general behaviors of students that could result in dropping out of school. In [16], a variety of Machine Learning algorithms were explored for development of Predictive Model based on incoming students data. In an effort to achieve models of students' performance and dropout, Naïve Bayes was explored in [14], K- Means in [7] and Decision trees in [22] and [13]. Furthermore, Predictive model to maximize students' retention and enrolment prediction was developed by [2], [19] and [24].

A review of predictive algorithms on student academic performance was carried out in [15]. The work seeks to determine the frequently used data mining algorithms in the domain. The research concluded that most of the research has been frequently done on same prediction algorithms, student variables and data mining tools. The C4.5 (J48) is often used by most researchers to predict the performance of the student, followed by ID3, CART and Naïve Bayes algorithms are frequently used by the researchers. The other algorithms such as ANN, Multilayer Perceptron and Bayelsian Classification are rarely used by the researchers.

Various application of Data mining in Education was investigated in [17]. The research includes analysis and visualization of data, predicting students' performance, enrolment management and grouping of students amongst others. The work discovered that Expected Maximization (EM), Apriori, C4.5 and Support Vector Machine (SVM) algorithms are more often used in Educational Data mining research.

Multiple data mining tasks was used to create qualitative predictive models for predicting the students' grades from a collected training dataset in [1]. The work explores multiple factors theoretically assumed to affect students' performance in higher education, and finds a qualitative model which best classifies and predicts the students' performance based on related personal and social factors. Four decision tree algorithms were implemented, namely, C4.5 decision tree, ID3 decision tree, CART decision Tree, and CHAID. It was found that the student's performance is not totally dependent on their academic efforts, but, on many other factors that have equal or greater influences.

The menace of student dropout is global. Therefore, researchers from all parts of the world have made innovative efforts at ameliorating its effects. [5], extends the modeling of students performance using data mining with a tutoring action plan to prevent academic dropout. Historical grades of students for different activities were used and the algorithm for the model is Logistic Regression. The approach reduced the dropout rate by 14%.

C4.5 and ID3 decision tree were adopted in [6] for constructing a predictive model of student dropout, characterizing students at the University Simón Bolívar, Miranda State, Venezuela in order to predict the probability that a student drop out of his/her an

academic program. WEKA was used for the two data mining techniques and for comparing their results.

In a University in India, [21] made effort to identify attributes from social-demographic, academic and institutional data from undergraduate students. The attributes were used to develop a model based on an improved ID3 decision tree. The traditional ID3 algorithm was improved using Renyi entropy, Information gain and Association function. The performance of improved algorithm measured on True Positive, False Positive, Recall, Precision and F-Measure were better that than the traditional ID3.

In 2017, a study with information from 2010-2015 and 2015-2016 generational Cohorts of Technological Institute of Zitacuaro – Mexico by [18]. The research adopts Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. The model is intended to accurately and promptly identify those students at risk of dropout so that universities can generate plans and strategies for retention. The model uses a DM methodology to learn models based on decision trees. The data collected include qualifications (performance) of students, and results of the EXANI II, which contemplates different personal, economic, social, cultural and academic factors.

III. MATERIALS AND METHODOLOGY

A machine learning project has a number of well-known processes, though it may not be linear. The five steps adopted for the dropout early warning model are Problem Definition, Data Preparation, Algorithms Evaluation, Results Improvement and Results Presentation.

3.1 Problem Definition

The second component of the Dropout monitoring and control system framework specifies the development of Dropout Alert and Control System. This early warning system is expected to beam the signal of dropout tendency to the drowning students and the management early enough for possible prevention. Considering the numerosity and diversity of possible factors that could be responsible for dropout, analytical modeling technique was considered inappropriate because of the complexity, hence, the choice of adaptive modeling. Consequently, the possible causes of dropout were identified to be either internal or external. The internal factors include behavioural issues, self-efficacy, academic challenges and health/biological issues. The external factors include early attainment of responsibility status, negative social-economic status, family configuration and family interaction. Others external factors are Peer influence and school non-conduciveness, uncoordinated records and weak support. [20] and [23].

3.2 Data Preparation

Data preparation is the process of determining the required attributes for a given Machine Learning problem. Mostly, live datasets are not naturally fit for the development of adaptive machine learning model. Largely because of noise, missing values and high diemensionality. It is therefore critical that the algorithms are fed with adequate data through proper data selection, data preprocessing and data transformation. These tasks which are considered to be about eighty percent (80%) of the total activities in Machine Learning models are handled as follows:

Data selection stage: The data selection provides the stage for determining which of the several attributes will be relevant for development of the model.This research is guided by the knowledge of the domain experts and the scope of the research. So, the databases considered in the students portal are the students' personal records, academic records, and entry grades in addition to the disciplinary records. The personal records table has fifty nine (59) direct attributes. Out of these numbers only ten (10) were selected. The ten purposely selected data include id, date_of_birth; place_of_birth; marital_status; sex;state_id; prog_id; level_of_entry; disability;sponsor_name.

In the academic records table, information about the students' course registration and academic performances are available. Data extracted from it include the first year Harmattan semesters' total number of unit (TNU) registered and the grade point average (GPA). Others are the cumulative total number of unit (CTNU), cumulative grade point average

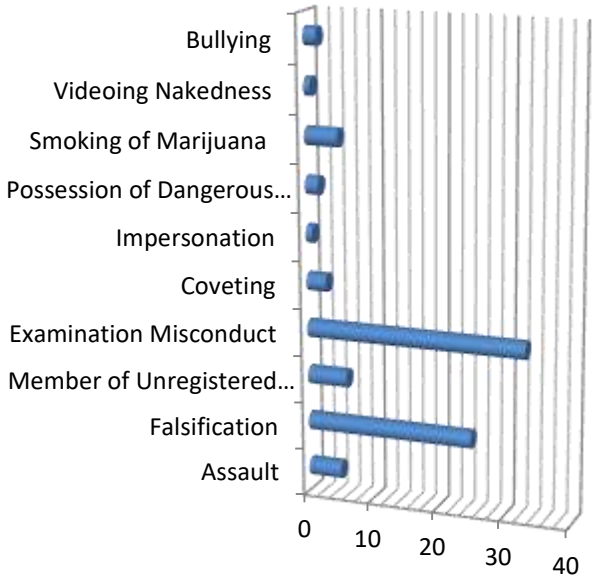
(CGPA) and total failed (outstanding) courses. The total number of students considered is eight hundred thirty four (834), out of which one hundred and eighty one (181) are Direct Entry (DE) Students. From this sample the number of withdrawal is forty four (44), thirteen of the withdrawn students are DE students. The Formular used for determining the Academic Status of the students is:

=IF(AND(K2<1.2,N2<1.2),"WITHDRAWAL",IF(AND(K2>1.2,N2<1.2),"PROBATION",IF(AND(K2<1.2,N2>1.2),"GOODSTANDING",IF(AND(K2>1.2,N2>1.2),"GOOD STANDING")))). (1)

The Senior Secondary School (O'level) grades obtained by the students in English Language, Mathematics and other three (3) relevant O'level subjects were extracted. The criteria that majorly influenced the sample records from the 834 records are the availability of O'level results, GPAs and CGPAs. The relations (tables) were processed to the third level of Normalisation (3NF) and necessary relationships were established.

The third category of data considered is the various types of misconducts in the Universities disciplinary records from 2013/2014 to 2016/2017. Eleven (11) unique items were identified. They include Assault (5), Falsification (25), Member of Unregistered Association (6), Examination Misconduct (32), Coveting (3), Impersonation (1), Possession of Dangerous Weapon (2), Smoking of Marijuana (5), Videoing Nakedness (1) and Bullying (2). The distribution is as shown in Table1.

Table 1: Types and distribution of offences - 2013/2014 to 2016/2017



The Data Preprocessing: The Data Preprocessing step is the next stage after the selection of appropriate dataset. The three major types of activities at this stage include; data formatting, data cleaning and data sampling. The choice of activities for a given problem is usually informed by the Machine Learning algorithms to be used and the sources of the dataset. The following are the description of the data preprocessing considered.

The dataset is from a relational database. For it to fit into the algorithms, it was formatted to a flat file structure. The resultant .csv file has twenty six (26) attributes. The cleaning of data has to do with the removal or fixing of missing data. There are several data instances that are incomplete in the dataset. To this end, one hundred and seventy-seven (177) records were removed for non-availability of O'level grades. From the remaining six hundred and fifty-seven (657) instances used the missed values across the attributes are as follows: Mathematics (2), English Language (2), Physics (Nil), Chemistry (3), Biology (1), No of sittings (6), JAMB Score/A'Grade (19), Year of Birth (43). The computed value technique was used for fixing the missing values. Each of the missing data was replaced by their respective mean values.

The population for the project is a state university in the North Central of Nigeria. Expectedly, there are far more data available than required for the dropouts model. So due to the need to limit the training times and other computational requirements, the Computer Science programme was selected for prototyping solutions before considering the whole university's dataset. The records of the whole nine sets (2009/2010 to 2017/2018) of students admitted in the Department of Computer Science were considered for building the model.

The Data Transformation: This final step of the data preparation stage in Machine Learning model is to convert the dataset to different forms as specifically required by the algorithms. Hence, the specific algorithm(s) and the knowledge of the problem domain influence this step. Three common data transformations are scaling, attribute decompositions and attribute aggregations. This step is also referred to as feature engineering. The following are the transformations applied.

Scaling: Admissions into Higher Educational Institution in Nigeria require satisfactory grades in Senior Secondary School Examination (SSCE, O'level) and Advanced level (A'level) Examination or Unified Tertiary Matriculation Examination (UTME) moderated by the Joint Admissions and Matriculations Board (JAMB). Grades from these examinations are not on the same scale. The UTME is rated over 400 points, IJMB results are rated over 15 points, and others like OND and NCE are nominal value. This necessitates transformation as shown in table 2.

Table 2: UTME and A'Levels grades scale

Exams Name	Point 1	Point 2	Point 3
UTME	< 200	200 – 299	300 -400
IJMB	1-5	6 – 10	11 – 15
OND and NCE	Pass, Merit and Lower Credit	Credit and Upper Credit	Distinction

The conventional O'level grades were also scaled to points as shown in table 3.

Table 3: SSSE - O'level grades scale

O'level Grades	A1	A2	A3	B2	B3	C2	C4	C5	C6
Points	9	8	7	6	5	4	3	2	1

Furthermore, the year of entry is transformed as shown in Table 4 to mean the cohort or set number.This is important because most of the classifiers would not accept strings data values.

Table 4: Year of entry / cohorts

17/18	16/17	15/16	14/15	13/14	12/13	11/12	10/11	09/10
9	8	7	6	5	4	3	2	1

Decomposition: The date of the birth attribute was split into the constituent parts of dd-mm-yy. The year (yy) component of it was used in the model. Similarly, the list of failed (outstanding) courses was also decomposed to its numbers for the Harmattan and Rain semesters.

Aggregation: The single features formulated by aggregating some basic features include, the total number of units registered (TNU), the total number of credit earned (TCE) and the grade point average (GPA). Similarly, the three variables previously (PTNU, PTCE and PGPA) and cumulatively (CTNU, CTCE and CGPA). Others are the numbers of carryovers for the Harmattan semester (HarmattanCO) and Rain semester (RainCO).

3.3 Algorithms Evaluation

The step adopted for determining the most appropriate algorithm for the model as proposed in [11] include Loading and Summarising dataset, Data Visualisation, Algorithms Evaluation and Making Predictions.

Loading and Summarising the dataset: Data summary and visualisation provide necessary insight and understanding of the historical dataset. This is essential because the discoveries play significant roles in the determination of choice of algorithms and types of

Table 5: Attributes Description

	entry_ session	entry_ level	birth_ year	sex	maths	english	physics
count	658	658	658	658	658	658	658
unique	10	3	20	3	10	9	9
top	8	1	1995	2	1	1	1
freq	192	495	101	482	192	290	221

	chemistry	biology	sittings	…	ptce	pgpa	ctnu
count	658	658	658	…	658	658	658
unique	9	10	3	…	152	220	28
top	1	1	1	…	46.4	2.3	42
freq	221	242	498	…	19	14	302

	ctce	cgpa	Harmattan -co	rainco	totalco	Mis- conduct	Acad- status
count	658	658	658	658	658	658	658
unique	226	246	11	11	15	3	3
top	103.2	1.94	0	0	0	0	GD
freq	12	10	236	328	168	655	613
[4 rows X 25 columns]							
acadstatus							
GD	613						
WD	44						
acadstatus	11						
dtype:	int64						

model. The 5 key Python SciPy libraries used are scipy, numpy, matplotlib, pandas and sklearn. These libraries and the spider API were installed seamlessly from Anaconda–https://www.continuum.io/downloads. The versions of the Python and libraries used for the experiment are:

Python: 3.6.3 |Anaconda, Inc.| (default, Oct 15 2017, 03:27:45) [MSC v.1900 64 bit (AMD64)]

scipy: 0.19.1 ; numpy: 1.13.3; matplotlib: 2.1.0; pandas: 0.20.3; sklearn: 0.19.1

The loading requires importing the modules, functions and objects. That was followed by loading the dataset into the Pandas tool.

The data dimension and the glance of the top-ten records revealed that the data set has twenty five (25) attributes and six hundred and fifty seven (657) instances. The rows show seven (7) students with 'Good Standing status' and three (3) withdrawals, thus revealing the dataset as a binary classification problem. Furthermore, the output of the Attributes description revealed the count, the unique, the top and the frequency value for each of the attributes. The data in Table 5 can further be described through the transformation scale in table 2, 3 and 4. The dataset for the attributes distribution included the label heading for each of the attributes, hence, the 658 shown as the count. The actual count of the instances is 657 and shown in all other reports. The output of the attributes description revealed that the largest set is the 2016/2017 set 192/657, the UTME students are the largest - 495/657 and the male enrolment in Computer Science is much higher than the female - 482/657.

Furthermore, the report showed that the most frequent O'level grades across Mathematics, English language and other sciences is C6 (point 1), the highest of this lowest credit is in English language - 298/657. It is also observed that most of the student had the five (5)

credits at a single sitting - 498/657. Finally, majority of the students are in good academic standing – 613/657.

From all indications, the historical data is an imbalance dataset on several attributes. This is usually the case with an educational dataset and considered as one of the key topics in machine learning community [12]. As such the efficacy and the necessity of Oversampling, Undersampling, Threshold-moving and Emsemble rebalancing techniques on the performance of linear and non-linear classifiers are recommended for further studies.

Data Visualisation: The basic idea about the data from the summaries can further be substantiated and extended by visualizations. The two common types of plots are the Univariate and Multivariate plots. The Univeariate plots are used to better understand each attribute, while the Multivariate plots are meant to better understand the relationships between attributes.

Univariate Plots: Whisker plot or Boxplots is a univariate plots that reveals the distribution of the input attributes. It summarizes the distribution of each attribute, drawing a line for the median (middle value) and a box around the 25th and 75th percentiles (the middle 50% of the data). The whiskers give an idea of the spread of the data and dots outside of the whiskers indicate candidate outlier values. The outlier values are 1.5 times greater than the size of spread of the middle 50% of the data. Figure 1reflect the distribution of the entry attributes.

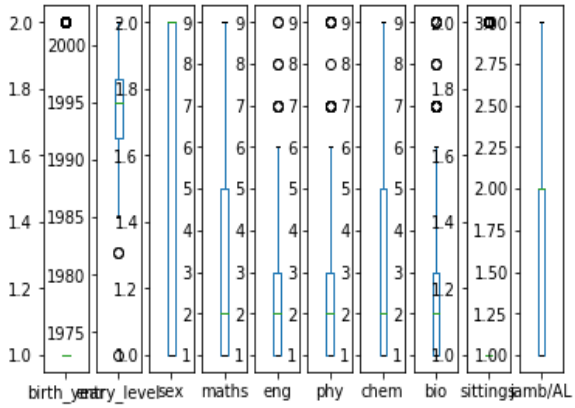


Figure 1: Box and Whisker Plot of Entry Attributes

In Figure 1, O'level grades A3, A2, and A1 are considered as outliers. Similarly, theoutliers are pointers to areas of attention during model training.

Another Univariate plot that gives an idea of the distribution is the histogram of each input variable. Histograms group data into bins and provide a count of the number of observations in each bin. From the shape of the bins as shown in Figure 2 and Figure 3, it is easier and quicker to get a feeling for whether an attribute is Gaussian', skewed or even has an exponential distribution. It also helps to immediately observe the possible outliers.

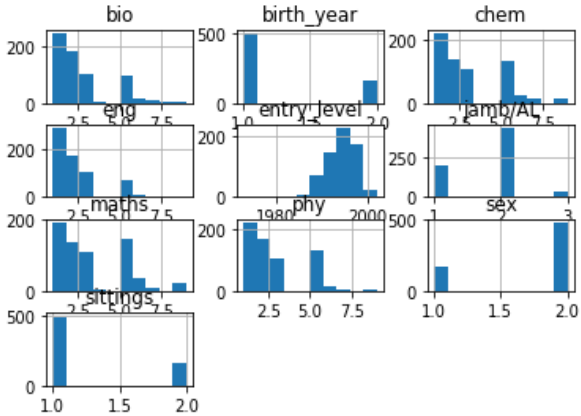


Figure 2: Histogram of Entry Input Variable

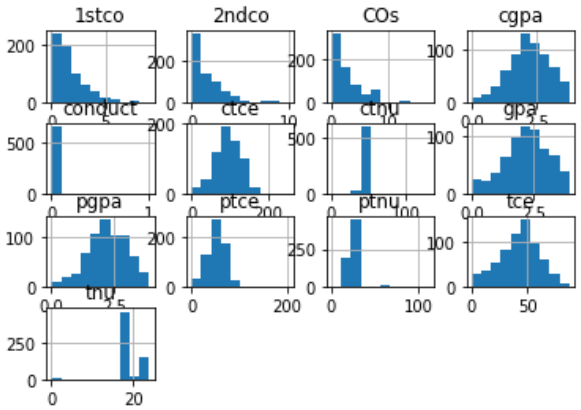


Figure 3: Histogram of University Academic Input Variable

Furthermore, it is instructive to note that some of the input variables such as cgpa, ctce, gpa, pgpa, tce and birth_year have perhaps a Gaussian distribution. This is

a useful guide for algorithms that can exploit this assumption. The other input variables like 1stCo, 2ndCo and Cos have perhaps exponential distribution.

Multivariate Plots: The multivariate plots provide means of observing the interactions between the variables. Figure 4shows the scatterplots of all pairs of attributes as dots in two dimensions, one axis for each attribute. Drawing all these scatterplots together is called a scatterplot matrix. This can be helpful in spotting the structured relationships between input variables. Figure 4 is a Scatter Plot of Entry Attributes.

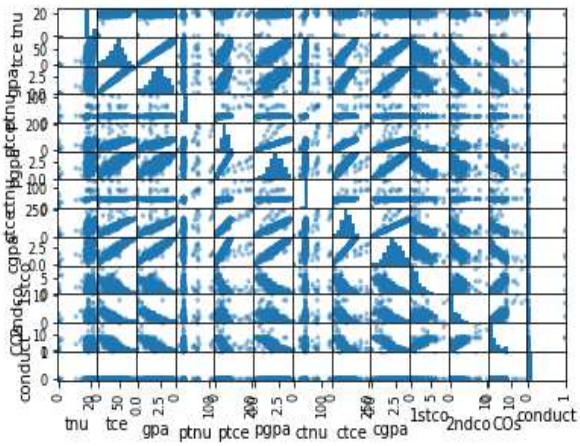


Figure 4: Scatter Plot University Academic Attributes

The scatterplot matrix is symmetrical. This is very useful when considering pair-wise relationships from different perspectives. Because there is little point drawing a scatterplot of each variable with itself, the diagonal shows histograms of each attribute. The diagonal grouping of some pairs of attributes suggests a high correlation and a predictable relationship. Attributes with structured relationships may also be correlated and thereby suggest candidates for removal from the dataset. The Figure 4 shows the correlations between the academic attributes, and provides a basis for trimming these attributes for better performance.

Algorithm Evaluation: Having gained insight into the dataset, the next stage is to create some models of the data and estimate their accuracy on unseen data. The steps involved require Separating out a validation dataset, Setting-up the test harness to use 10-fold cross-validation, Building 6 different models to predict dropout and selecting the best model.

Create a Validation Dataset: Determining the goodness of predictive models could be by statistical methods or by a concrete estimate of the accuracy of the best model on unseen data. The latter approach is known as the validation method. The method requires that certain percentage of the dataset be held back, such that that the algorithms will not get to interact with the excluded data. The excluded data would eventually be used to determine the accuracy of the best model. This work follows the common practice of splitting the loaded dataset into two, 80% for training the models and 20% as a validation dataset.

Set-up the test harness to use 10-fold cross-validation: The next step is to set-up the test harness to use 10-fold cross-validation to estimate accuracy. This will split the dataset into 10 parts, train on 9 and test on 1 and repeat for all combinations of train-test splits. Model accuracy is the metric used to evaluate models.

Build 6 different models to predict dropouts: Experimental approach provides an adequate method for determining the most appropriate algorithms and configurations for adaptive models. Consequently, the six (6) algorithms explored are Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Gaussian Naive Bayes (NB) and Support Vector Machines (SVM).

This collection is a good mixture of simple linear (LR and LDA), nonlinear (KNN, CART, NB and SVM) algorithms. To ensure that the evaluation of each algorithm is performed using exactly the same data splits, the random number seed is reset before each run. This ensures that the results are directly comparable.

Selection of the best model: The best model was determined experimentally. The six algorithms were tested with the full attributes and the two trimmed attributes (the one based on entry attributes and the other academic attributes). The performance metrics used are Accuracy, Precision, Recall and F1-Score.

Accuracy: Classification Accuracy is the ratio of number of correct predictions to the total number of input samples. It is given as:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{TP + FP + FN + TN} \quad (2)$$

It works well only if there are equal number of samples belonging to each class.

Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (3)$$

Recall: It is the number of correct positive results divided by the number of ***all*** relevant samples (all samples that should have been identified as positive).

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (4)$$

F1 Score is the Harmonic Mean between precision and recall. It tries to find the balance between precision and recall. The range for F1 Score is [0, 1]. It tells how precise the classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

High precision but lower recall, gives an extremely accuracy, but it then misses a large number of instances that are difficult to classify. In this instance, it is observed that the greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as:

$$F1 = 2 * \frac{1}{1/_{Precision} + 1/_{Recall}} \quad (5)$$

Table 6: Result of Model Performance on Full Attributes Entry Attributes and First Year Results

Algorithms	Full Attributes	Entry Attributes	First Year Results	Ave. Performance
LR	0.986720 (0.012116)	0.929572 (0.026967)	0.988607 (0.012560)	0.9683 (0.017215)
LDA	0.960160 (0.034172)	0.927649 (0.030552)	0.969594 (0.029530)	0.952468 (0.031418)
KNN	0.990493 (0.009507)	0.923803 (0.024165)	0.990493 (0.009507)	0.968263 (0.014393)
CART	0.986756 (0.014751)	0.878157 (0.053873)	0.990530 (0.012684)	0.951815 (0.027103)
NB	0.942925 (0.030383)	0.929572 (0.026967)	0.937192 (0.030386)	0.936563 (0.029246)
SVM	0.944811 (0.024683)	0.931459 (0.027207)	0.956205 (0.025488)	0.944159 (0.025793)

The Mean Accuracies of the six algorithms are as shown in Table 6. Also shown in brackets are the Standard deviations which provide insight on the precisions of the algorithms. From the mean figures, it could be realized that the most accurate algorithm for the full dataset is K-Nearest Neighbour (KNN), follow by Classification and Regression Trees (CART) and the least is Gaussian Naïve Bayes (NB). The most accurate on Entry grades is Support Vector Machine (SVM), jointly followed Linear Regression (NR) and NB while Classification and Regression Trees (CART) is the least. Furthermore, the Classification and Regression Trees (CART) gives the best accuracy on the University first-year grade, followed by KNN while Naïve Bayes (NB) is the least. KNN provides the least Standard deviations which are indications of highest precision across the three categories. The Standard deviation (precision) figures of KNN in addition to being the best and second best on full data and first-year data respectively makes it the best of all the six and the choice for dropout modeling. These performances on the entry attributesare further illustrated in Figures 5. The superior performance of the KNN is clearer in the average performance

accuracy of 0.968263 which is the highest and 0.014393 standard deviation which is the lowest.

The best of the algorithms identified was K-Nearest Neighbour (KNN). It was therefore used to test the classification accuracy of the model.

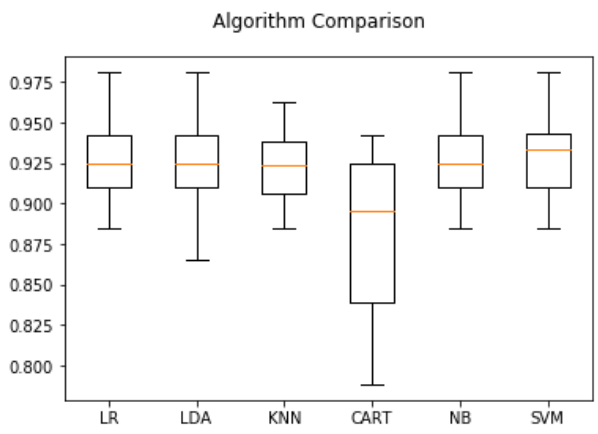


Figure 5: Comparison of Algorithms on Entry Attributes

The result of the *classification_report()* is as shown in Table 7.

Table 7: Classification Report of the KNN based Dropout Model

Attributes Categories	Classes	Precision	Recall	F1-score	Support
Full Attributes	GD	0.99	0.99	0.99	125
	WD	0.86	0.86	0.86	7
	Avg / Total	0.98	0.98	0.98	132
Entry Attributes	GD	0.95	1.00	0.97	125
	WD	0.00	0.00	0.00	7
	Avg / Total	0.90	0.95	0.92	132
1stYear Academic Attributes	GD	0.99	0.99	0.99	125
	WD	0.86	0.86	0.86	7
	Avg / Total	0.98	0.98	0.98	132
GD: Good Standing; WD: Withdrawn					

The classification_report() function displays the precision, recall, F1-score and support for each class. The values show good precision, recall and F1-score of 0.98 for the full attributes and first-year results. The performance evaluation for the Entry Attributes is low for Good-Standing and extremely low on the Withdrawal class. This calls for further studies to see the effect of data transformation and /or data rebalancing on the model. The Performance evaluation of academic attributes compared favourably with the classification obtained with the full attributes. The implication of this fact, among others, is that the academic factor which is a readily available data is enough to trigger management intervention for preventing the affected student from dropping out even in the absence of other details. Furthermore, it could spark a purposeful investigation into other causative factors for remediation.

CONCLUSION

This research developed an adaptive modelbased on the earlier proposed framework for dropout warning systemaimed to provide an early signal on the students with the propensity to dropout.The simulation results revealed that the performance evaluation of academic attributes compared favourably with the classification obtained with the full attributes. The implication of this fact, among others, is that the academic factor which is a readily available data is enough to trigger management intervention for preventing the affected student from dropping out even in the absence of other details. Furthermore, it could spark a purposeful investigation into other causative factors for remediation.

FUTURE WORK

Future efforts would focus on exploring the effects of data rebalancing and feature selection on the model configuration and performance. Also, as a build on to the realization of the Adaptive Dropout Model, further work would be the finalization of the model and implementation as application software over an expanded scope.

ACKNOWLEDGEMENT

Tertiary Education Trust Fund (TETFUND) provides support for the project under the Institutional Based Research (IBR) Grants.

REFERENCES

[1] Amjad-Abu S. (2016). Educational Data Mining and Students' Performance Prediction. International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, pp.212-220

[2] Antons C, Maltz E., Expanding the role of institutional research at small private universities: a case study in enrollment management using data mining. *New Direction of Instructional Research*, (2006), 131:69–81.

[3] Baepler P, Murdoch C.J., Academic analytics and data mining in higher education. *International Journal of School Teach Learn*, (2010), 4(2): 1–9.

[4] Baker R., Yacef K., The state of educational data mining in 2009: a review and future visions. *Journal of Educational Data Mining*, (2009), 1(1): 3–17.

[5] Concepción B., María L.C., David P., Juan A.L., David L., María A.M., Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering*, (2018), 66, 541-556, https://www.sciencedirect.com/science/article/pii/S0045790617305220.

[6] Diana H. , Yegny A. and Edwin B. , Student Dropout Predictive Model Using Data Mining Techniques. *IEEE Latin America Transactions* , (2015), 13(9), https://ieeexplore.ieee.org/document/7350068.

[7] Erdogan S.Z and Timor M. , A data mining application in a student database. *Journal of Aeronautic Space Technolpgy*, (2005), 2(2):53–57.

[8] Iam-On, N. and Boongeon, T., Generating descriptive model for students' dropout: a review of clustering approach. *Human-centric Computing and Information Sciences*, (2017), 7(1), https://hcis-journal.springeropen.com.

[9] Isiaka R.M., A Framework for Dropout Monitoring and Management System. *Computing, Information Systems, Development Informatics & Allied Research Journal*, (2017), 8(1), - www.cisdijournal.net.

[10] Jadric M., GaračaŽ.,Ćukušić M., Student dropout analysis with application of data mining methods, *Management*, (2010), 15 (1), 31- 46, https://hrcak.srce.hr/file/81744.

[11] Jason B., The Missing Roadmap to Self-Study Machine Learning, (2014), Retrieved from https://machinelearningmastery.com

[12] Jianhong Y., Suqing H., Classifying Imbalanced Data Sets by a Novel RE-Sample andCost-Sensitive Stacked Generalization Method, *Mathematical Problems in Engineering*, (2018), Article ID 5036710, 1-13, https://doi.org/10.1155/2018/5036710

[13] Kabra R.R and Bichkar R.S., Performance prediction of engineering students using decision trees. *Internal Journal Computer Applications*, (2011), 36(11):8–12.

[14] Kotsiantis S, Pierrakeas C, Pintelas P., Prediction of student's performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, (2004), 18(5): 411–426.

[15] Kumar D., Selvam P., Kumar S., (2018). Review on Prediction Algorithms in Educational Data Mining. International Journal of Pure and Applied Mathematics. Volume 118 No. 8, pp. 531-537

[16] Lin S.H., Data mining for student retention management. *Journal of Computing Science Collection*, (2012), 27(4): 92–99.

[17] Nithya P., Umamaheswari B., and Umadevi A. (2016). A Survey on Educational Data Mining in Field of Education. International Journal of Advanced Research in Computer Engineering & Technology. Volume 5. No. 1. Pp.69-78

[18] Noel E. R, Carlos L. , Oscar M., Brian A.S. Modeling Students' Dropout in Mexican Universities. *Research in Computing Science*, (2017), 163–175, ISSN 1870-4069.

[19] Ramaswami M, Bhaskaran R., A CHAID based performance prediction model in educational data mining. *Internal Journal Computer Science*, (2010), 7(1):10–18.

[20] Shahrina M.N., Wan-Ahmad W.F., Nayan Y., Yahya N., Abdullah A., Abdul Rahman A., Ismail N, and Ahmad N.A., A Conceptual Framework in Examining the Contributing Factors to Low Academic Achievement: Self-Efficacy, Cognitive Ability, Support System and Socio-Economic. *International Conference On Management, Social Sciences And Humanities*, (2012), Available from: https://www.researchgate.net/publication/259470497.

[21] SubithaSivakumar, SivakumarVenkataraman and RajalakshmiSelvaraj, Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree, *Indian Journal of Science and Technology*, (2016), 9(4), www.indjst.org/index.php/indjst/article/download/87032/66961, DOI: 10.17485/ijst/2016/v9i4/87032.

[22] Sung-Hyuk C, Tappert C., Constructing binary decision trees using genetic algorithms. *Journal of Pattern Recognition Research*, (2009), 1:1–13.

[23] Wallace C.M., A High School Dropout Prevention Program for At-Risk Students. *Walden University Scholarworks*, (2016), Available from http://scholarworks.waldenu.edu/dissertations.

[24] Yu C, Gangi S.D, Jannasch-Pennell A., Kaprolet C., A data mining approach for identifying predictors of student retention from sophomore to junior year, *Journal of Data Science*, (2010), 8:307–325.