A model for the identification of students at risk of dropout at a university of technology

Roderick Lottering
Department of Computer Science
Tshwane University of Technology
Soshanguve, Pretoria, South Africa
lottering.roy@gmail.com

Robert Hans
Department of Computer Science
Tshwane University of Technology
Soshanguve, Pretoria, South Africa
HansR@tut.ac.za

Manoj Lall
Department of Computer Science
Tshwane University of Technology
Soshanguve, Pretoria, South Africa
LallM@tut.ac.za

***Abstract*— Development has been seen in the advanced education segment in South Africa. With this development, an expansion in the dropout rate is noticed. This study explores the adequacy of dimensional decrease and concentrates the significant data covered up in the student information for the identification of students in danger of dropout. This study depends on educational data mining techniques and makes forecasts of dropout goal of understudies from courses. In the test, the researchers show promising outcomes with information from the recognition courses of a University of Technology.**

Keywords—educational data mining, feature selection, prediction, student dropout, learner analytics

I. INTRODUCTION

Student numbers in higher education in South Africa grew by 32.8% from 2006 to 2015 [17]. Furthermore, the Department of Higher Education [9] reported that a dropout rate of 17.1% was observed among cohort enrolments for a three-year National diploma through contact mode nationally for the 2015 academic year. This dropout rate is expected to double over ten years [9]. Furthermore, the Council of Higher Education [7] reported in a 2016 study that students are taking up to three years longer than the regulation time to complete their studies among cohorts for the diploma studies. The dropout rate increased from 35% to 51% over the period 2013 to 2016 [7].

Large datasets exist about students' education foundations with information, for example, their learning and environments in which they study. Universities start to see how to use this asset to improve the instructive experience for students [8]. Govindarajan, et al, [10] referenced that predictive techniques could be utilized to discover the 'in danger' students who face challenges in the learning procedure that lead to course disappointment circumstances or pull back from the course or organization. The word 'in danger' normally identified with the presentation to some danger or harm. In the context of this study, this term alludes to students who may drop out of a course or who don't meet the passing criteria [31].

Learner Analytics (LA) provides the necessary algorithms that can be applied to educational data already collected by institutions of higher learning [13,33,35]. This data can be used to discover patterns that lead to an identification of students at academic risk [12].

As a link has been established between dropouts and students at risk [31], this study aims to use Educational Data Mining (EDM) techniques to identify students at risk of dropout at the Information Communication and Technology (ICT) faculty of a University of Technology (UoT). With this in mind, the following research question has been formulated:

How can a predictive model be designed and developed to identify students at risk of dropping out at a University of Technology?

With the identification of students in danger of dropping out, improved student retention and risk attrition could be some of the advantages of LA and EDM [33]. The motivation behind this research project is to assemble a predictive model to recognize students in danger of drop out, with the use of EDM. To achieve this, factor analysis and educational mining techniques were applied to student data sourced from the ICT faculty of a university. The EDM output including accuracy, sensitivity, specificity and the F-Measure were compared to express their predictive power.

The rest of this report is structured as follows: The next Section reviews theory and practice found on the identification of students at risk of dropping out, documented in the literature. Section III elaborates on the design and methodology used in this study. Section IV presents the research results and analysis. With Section V, conclusions, as well as limitations of this study, are presented.

II. LITERATURE REVIEW

Digital data is the currency at large today and present in all organizations that embrace information technology. Data in the world increase rapidly therefore the capacity to analyze large data sets – big data – became a competitive force, underpinned by new waves of productivity, growth and innovation. Big data is a phrase that describes structured, semi-structured and unstructured data that is made available with electronic traces that everyone leaves behind when working online [13]. Nicole [23] noted, "Big data refers to the analysis of data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze". Every time a student interacts with their university through the various platforms, a digital footprint is left [31]. Lonto, van der Walt and Conradie [19] indicate that "Higher education institutions collect big data which include (i) course selection and registration, (ii) financial and work arrangements, (iii) class participation and study groups activity, (iv) online resource usage and (v) textbook purchase. This big data trapped in institutional systems could be used to help students, managers and administrators make better and informed decisions." "The goal of learner analytics is to uncover hidden patterns in educational data and use those patterns to attain a better understanding of the educational process, assess student learning and make predictions on performance" [12].

978-1-7281-6770-1/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: CALIFORNIA INSTITUTE OF TECHNOLOGY. Downloaded on September 26,2020 at 11:38:14 UTC from IEEE Xplore. Restrictions apply.

A. *Data Mining*

Many reasons as indicated by Dietz-Uhler and Hurn [8] spurs an enthusiasm for LA. Some of these reasons for the increased interest in LA are the general trend of increased accountability in all levels of education. LA takes into account the assortment of information from activities performed by students on different stages where they connect with course content. Data mining can be applied to discover fascinating and unforeseen connections among properties of students, instructing and techniques and evaluations as detailed by [4] and [31]. "Data mining strategies are viewed as better measurable techniques compared to statistical methods as the data size may be huge with the challenge of processing large datasets using statistical methods [11].

"The main function of data mining is applying various methods and algorithms to discover and extract patterns of stored data. Data Mining (DM) and knowledge discovery applications have a rich focus due to its significance in decision making and it has become an essential component of various organisations" [3]. DM is a computational procedure of finding designs in large informational collections. It is a procedure that is interdisciplinary, grasping man-made brainpower, statistics, EDM and database frameworks. Data Mining is a phase that is alluded to as the investigation stage and its errands are grouping, inconsistency discovery, clustering and affiliation rule mining [24].

Classification is the most frequently applied data mining technique, which is based on a set of pre-classified examples to develop a model that can classify the population of records at large [3]. Classification techniques in data mining can process large quantities of data. It tends to be utilized to predict categorical class labels and classifies data based on training sets and class labels and used for classifying newly available data [21]. The generation of data stored in databases, data warehouses and datasets might be influenced by unusual behaviours creating anomalous data. The outlier analysis data mining task aims to discover abnormal patterns in datasets or deviation in regards to the normal behaviour [18]. Outlier analysis is also referred to as anomaly detection, anomaly mining or deviation detection. A common technique for data analysis used in the fields of pattern recognition, information retrieval, bioinformatics, machine learning and image analysis is Clustering. This mining technique can be achieved by algorithms that differ in similarities required between elements of a cluster and how to find the elements of the clusters efficiently [1]. Association is all about finding frequent patterns, associations, correlations or causal structures among sets of items or objects in transactional databases, relational databases, and other information repositories [33]. Association Analysis helps in finding concealed patterns from large available data sets. It aims to discover correlations and patterns among data of interest [31].

In clustering, groups of objects are assigned into clusters that are discovered from the data. Objects in the same cluster are more similar to each other than the objects in other clusters [27]. Srilekshmi, et al., [34] define Association Analysis as "helps in discovering hidden patterns from large available data sets". It aims to discover interesting correlations and frequent patterns among data of interest. Since our goal is to classify students' performance into the predefined classes - "At-Risk" and "Not At Risk", clustering and association are not suitable choices and so the researchers have used classification algorithms and discarded the other methods.

B. *Factors to identify a student at risk of dropout*

Attributes, factors, features and variables are used interchangeably in this report. Gulati [11], concluded that many factors such as demographic factors, socio-economic factors and family factors are related to students dropping out from courses. This drop out means stop schooling for valid reasons and disappointment with their social system and examination results [15].

Kabakchieva, [14] in her study of predicting student performance included: (i) student background factors (gender, birth year, birthplace and living place in the country), (ii) previous academic performance (previous type of education, profile and place of previous education and total score of previous education) and (iii) current academic factors (university admittance year, admittance exam and achieved score, university speciality/direction, current semester and total university score) in a dataset. The final dataset used for the project implementation contains 10330 instances (539 in the "excellent" category, 4 336 in the "very good" category, 4 543 in the "good" category, 347 in the "average" category, and 564 in the "bad" category), each described with fourteen attributes (1 output and 13 input variables), nominal and numeric. Gender and age were attributed that related to the student personal data.

Thomas, [39] investigated college completion for undergraduates by applying a structural equation model to help understand the relationship among factors and how they contribute to students' College Completion Intention (CCI). Data collected through two questionnaires; one focusing on student demographics and another 52 Likert-type questionnaire that measured perception of the variables of the study. The data sourced was used to test various formulated hypotheses.

Srilekshmi, et al, [35] used a dataset of 16 courses with 20 attributes from the universities database in their study to identify students at-risk. The selected attributes held personal information as well as academic records of students including their learning background. The final dataset included six variables.

Steward, Lim & Kim [37] reported that factors that determine student persistence include pre-college experience, financial aid, college academic performance and student demographics including gender. This study had a dataset of 27 000 records which has sourced from the university database. Bergin et al, [5] define background factors as "Previous academic experience, grades achieved in second level exit examinations; previous experience of computer applications, game playing, internet usage and programming; the number of hours spent studying and working at a part-time job etc".

"Several studies indicate that one of the important factors of students' dropout rate is the subject studied at University as well as the secondary school grades. Indeed, the dropout rate is higher among students in engineering disciplines, and among students with relatively low levels of prior qualifications " [25].

Paura & Arhipova [28] found that students with a higher proportion of drop out are those enrolled in the faculty of ICT compared to other faculties. A low Maths GPA score elicits poor knowledge of the subject and this causes students' dropout rates in engineering sciences.

Authorized licensed use limited to: CALIFORNIA INSTITUTE OF TECHNOLOGY. Downloaded on September 26,2020 at 11:38:14 UTC from IEEE Xplore. Restrictions apply.

C. *Algorithms that predict students at risk of dropout*

Bergin, et al., [5] reported, "Identifying struggling students at an early stage is not easy as introductory programming modules often have a high student to lecturer ratio (100:1 or greater) and early assessment may not be a reliable indicator of overall performance". Interventions to prevent struggling students from failing or withdrawing from the course may often be too late by the time feedback is available. Their study included factors that influence programming success by applying five classifiers for determining programming performance including Logistic Regression (LR), K-Nearest Neighbor (KNN), C4.5, Naïve Bayes (NB) and SVM using Sequential Minimal Optimization (SMO). Three measurement techniques such as overall classifier accuracy, specificity and sensitivity were used in this study.

Liang & Zheng, [17] focused on student performance in an online platform by analyzing the behaviour logs of Users in Massive Open Online Courses (MOOCs). Metadata about courses, student's course enrollment records and the user behaviour logs were extracted from their online platform for analysis. Three classification features were defined in the analysis as UserFeature, CourseFeature and EnrollmentFeature with various variables associated with these features. Their data sample included 39 courses data, with each course containing 40 days of user behaviour logs for more than 20 000 students. Supervised classifiers such as Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), decision-making tree was applied to this problem. Training and tuning of these models resulted in achieving the best result of 88% accuracy for the Gradient Boosting Decision Tree (GBDT).

Aulck, et al., [2] analyzed student dropout in a large, heterogeneous population. The data contained the demographic information, pre-college entry information and complete transcript records for all students at the university. Their focus was on matriculated undergraduate students at their main campus who first enrolled over a selected period. The selection resulted in a population of 69 116 students over several years. Those who did not complete at least one undergraduate degree over a defined period were defined as students who dropped out. Three classifiers (LR, KNN and Random Forests) was applied to predict the binary dropout variable. The Receiver Operating Characteristic (ROC) curves for each of the models resulted in 66.59% accuracy for the logistic regression model. Among the strongest individual predictors for student retention was GPA in Mathematics, English, Chemistry and Psychology classes. Some of the conclusions of their study were that regularized logistic regression provided the strongest predictions for the dataset.

Algorithms had been identified in the above-mentioned literature for the application of predictive modelling in LA. From these studies, authors propose the following algorithms to predict student performance including (i) Logistic Regression, (ii) Decision Trees (DT), (iii) K-Nearest Neighbor (KNN), (iv) Naïve Bayes (NB) and (v) Support Vector Machines (SVM) to mention a those that feature prominently.

III. METHODOLOGY

A. *Data gathering and cleaning*

To conduct this research, ethical clearance and permission was received through internal processes to harvest data from the university database of 4417 full-time students who enrolled for the normal diploma courses at the ICT faculty of a South African university of technology. Table 1 provides a breakdown of the 9 normal diploma offerings for fulltime students between 2013 and 2017. These courses consist of three academic years with 24 courses each. Three thousand credits are allocated to a course.

The collected information consists of the final grade of the different courses, summarized in Table 1, taken by the students during their academic years. Various variables are extracted from the university database while some variables are derived from the Extraction, Transformational and Load (ETL) process (Figure 1) to enrich the dataset. All usable variables have been extracted into a final dataset with possible values described in Table 2 after data consolidation.

Table 1 – Normal diploma course and enrollments

Course	Academic year					Total
	2013	2014	2015	2016	2017	
Business Applications	1	25	50	44	198	**318**
Intelligent Industrial Systems	0	8	13	39	108	**168**
Communication Networks	0	3	44	95	216	**358**
Technical Applications	0	17	40	70	202	**329**
Support Services	0	10	25	39	159	**233**
Software development	0	24	63	147	614	**848**
Information Technology	160	452	404	396	539	**1951**
Web and Applications Development	0	4	4	28	30	**66**
Multimedia	0	15	18	38	75	**146**
Total	**161**	**558**	**661**	**896**	**2141**	**4417**

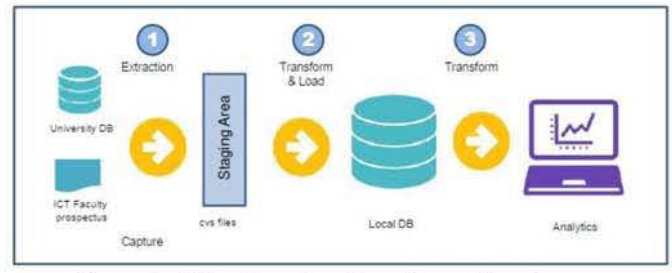


Figure 1 - Extraction, Transformation and Load process

Table 2 - Dataset with attributes and values

ATTRIBUTE	VALUES
1. ACCOMODATION	Y,N
2. AGE	Min 17, Max 48
3. CALENDER_YEAR	2013, 2014, 2015, 2016, 2017

Authorized licensed use limited to: CALIFORNIA INSTITUTE OF TECHNOLOGY. Downloaded on September 26,2020 at 11:38:14 UTC from IEEE Xplore. Restrictions apply.

ATTRIBUTE	VALUES
4. CREDITS_COLLECTED	Min 0, Max 3480
5. DISABILITY	Y,N
6. FINANCIAL_AID	Y,N
7. GENDER	M,F
8. HOME_LANGAUGE	AFRI,ENGL,ISIN, ISIX,ISIZ,OTHR, SESO,SESS,SETS, SISW,TSHI,UNKN, XITS
9. MODULES_COMPLETED	Min 0, Max 55
10. MODULES_REPEATED	Min 0, Max 22
11. MODULES_FIRST_TIME	Min 0, Max 24
12. PERSISTENCE	Min 0, Max 70
13. PREVIOUS_YEAR_ACTIVITY	S,NS
14. QUALIFICATION	NDIB12,NDIBF1, NDII12, NDIIF1, NDIK12,NDIKF1, NDIL12,NDILF1, NDIP12,NDIPF1, NDIS12,NDISF1, NDIT12,NDITF1, NDUI12,NDUIF1, NDIW12,NDIWF1
15. QUALIFICATION_PERFORMANCE	Min 1000, Max 3000
16. RISK_RATIO	Min 0 , Max 100
17. SEMESTERS	Min 1, Max 8
18. YEARS_IN_SYSTEM	Min 1, Max 7
19. FINAL_DECISION	AT_RISK, NOT_AT_RISK

All the data was stored in a relational database for data transition and manipulation purposes. The ETL process resulted in a dataset that did not have missing values, indicated with *NaNs* (Not a Number). All non-numerical values were converted into categorical values.

Data cleaning. The data was cleaned according to the following categories:

- *Missing values* – No missing values were detected in the dataset through a variance analysis done on the original dataset.
- *Outlier detection* – The original data set resulted in some outliers A total of 1815 student, identified as outliers, were removed from the original data set. The outliers corresponded to rare cases that would bias the results of the models. Table 3 provides the data set with all the outliers removed.

B. *Dropout prediction*

This study attempts to answer the question: is it possible to predict if a student is at risk of dropout of an ICT course given the information on current performance from the university database? As stated by the Council of Higher Education [7], an increase in dropout was observed between 2013 to 2016. Thus the researchers consider suitable to constrain this research to study students who did not complete their studies during the period 2013 to 2017. Table 3 shows the students who are still studying towards a diploma qualification through the normal during the indicated academic years.

RISK_RATIO is a derived value that is based on the following calculation presented in equation one with values selected from Table 4.

Table 3 - Normal diploma data set without outliers

Course	Academic year					Total
	2013	2014	2015	2016	2017	
Business Applications	1	9	13	13	86	**122**
Intelligent Industrial Systems	0	7	7	18	73	**105**
Communication Networks	0	1	17	19	122	**159**
Technical Applications	0	8	12	16	64	**100**
Support Services	0	8	6	19	78	**111**
Software development	0	8	21	39	253	**321**
Information Technology	101	374	301	286	416	**1478**
Web and Applications Development	0	0	0	0	0	**0**
Multimedia	0	6	6	7	32	**51**
Total	**102**	**421**	**383**	**417**	**1124**	**2447**

Table 4 - Academic Year Credit value

Academic Year (AY)	Required credits
Year 1	1000
Year 2	2250
Year 3	3000

$$\textbf{RISK\_RATIO} = \textbf{CC} / \textbf{AY} \qquad (1)$$

Where CC (CREDITS_COLLECTED) is the total number of credits a student collected to date on modules passed and AY is the Academic Year with a value based on Table 4.

The value assigned to the FINAL_DECISION variable is as follow:

Criteria	FINAL_DECISION
If RISK_RATIO >= 0.50	**"Not_At_Risk"**
If RISK_RATIO < 0.50	**"At_Risk"**

Authorized licensed use limited to: CALIFORNIA INSTITUTE OF TECHNOLOGY. Downloaded on September 26,2020 at 11:38:14 UTC from IEEE Xplore. Restrictions apply.

Equation one applies to all the students in the data set since the researchers want to enrich our data set further with a risk value assigned to the FINAL_DECISON variable of every student in the data set.

The dropout problem in Table 5, is imbalanced. This binary classification problem can be solved by the following two-step procedure:

Step 1: Feature vector and data pre-processing. Each student in the data set describes an *n*-dimensional vector with grades of all courses a student took. For the normal national diploma $n = 24$ for all the academic years. Students in the "Not_At_Risk" category are 578 compared to the 1869 students in the "At_Risk" category. The "At_Risk" category is the largest representation in this dataset. To properly train the classifiers the researchers used under-sampling of the "At_Risk" category to an equal sample size of the :Not_At_Risk" category to balance the dataset.

Step 2: Classification. Since this is a binary classifier in FINAL_DECISION that contain values "At-Risk" or "Not At Risk", the researchers trained the following suitable classifiers for such cases: *Decision Trees (C5.0)* [14], *Support Vector Machine* [22], *Naïve Bayes* [14] and *Nearest Neighbor* [5] using the feature vector of the training set samples. These classifiers were selected since they featured dominantly in the literature to solve a binary classification problem of predicting students at risk of dropout. *Random Forest* [23] was introduced as an additional classifier to test its ensemble properties. A brief explanation of each model is as follow:

DT algorithms describe the relationship between attributes and the relative importance of attributes. This is a "non-parametric classifier which works by partitioning the feature vector space one attribute at a time; interior nodes in the tree correspond to partitioning rules, and leaf nodes correspond to class labels. A feature vector x is classified by walking the tree starting from the root, using the partitioning rule at each node to decide which branch to take until a leaf node is encountered. The class at the leaf node is the result of the classification."

"SVM is an algorithm based on the idea of separating data using hyper-planes. It considers a set of *n*-dimensional feature vectors as points in the *n*-dimensional real euclidean space. It supposes that each point is associated with one class (0 or 1) and solves the problem of separating the points from each class by finding the hyperplane which is at the largest distance from both points of class 0 and points of class 1" [32].

"NB is a conditional probability model based in Bayes' theorem. Given an *n*-dimensional feature vector $(x_1,\ldots..x_n)$ and a classification class C, the algorithm computes $p(C|x_1,\ldots..x_n)$ using the Bayes' theorem. Independence between features is assumed" [32]. Combining this with a decision rule, the classification $\hat{y}$ for a feature vector $(x_1,\ldots..x_n)$ is done as follows in equation two:

$$\hat{y} = \underset{j\epsilon(1,2)}{\arg\max}\, p(C_j) \prod_{i=1}^{n} p(x_i|C_j) \qquad (2)$$

Table 5 - Dropout

	Not_At_Risk	At_Risk	Total
All courses	578	1869	**2447**

Nearest neighbor (NN) rule distinguishes the classification of unknown data point based on its closest neighbor whose class is already known. K-Nearest Neighbor (KNN) in which the nearest neighbor is computed based on the estimation of k indicates how many nearest neighbors are to be considered to characterize the class of a sample data point. "This type of learning is 'lazy' as it defers generalization until the classification stage. The NN algorithm is based on the principle that the properties of any particular instance are likely to be similar to those instances within its neighborhood" [32]. Each new instance is compared with existing ones using a distance metric and the new instance is classified based on the majority class of the nearest K neighbors.

Random Forest (RF) classifiers are ensemble learning techniques that work by constructing multiple Decision Trees and outputs are the mode of the classes of the individual trees. This model is trained using Feature bagging [32].

C. *Implementation*

All the analysis, data manipulation and data visualization used in the experiments are implemented in R [30]. Table 6 provides a list of R libraries (packages) that are used for the implementation of this research. For data pre-processing, the researchers used MS Access (Relational Database Management System), MS Excel with Visual Basic for Applications (VBA) embedded.

D. *Dimensionality reduction*

"Modern data sets are often described with far too many variables for practical model building. Usually, most of these variables are irrelevant to the classification, and their relevance is not known in advance. There are several disadvantages of dealing with overlarge feature sets" [15].

One of the most popular techniques to remove noisy (i.e. irrelevant) and redundant features is dimensional reduction. This reduction technique can be categorized mainly into feature extraction and feature selection. Feature extraction methods create new features by transforming original features into distinct feature spaces. Principle Component Analysis (PCA) is a popular feature extraction method. Feature selection methods select subsets from the original data sets according to criteria of feature importance. Feature selection maintains the physical properties of the original features compared to feature extraction and has been widely used in time series analysis and pattern classification [39].

"Feature extraction and feature selection methods are capable of improving learning performance, lowering computational complexity, building better generalizable models, and decreasing required storage" [32].

Authorized licensed use limited to: CALIFORNIA INSTITUTE OF TECHNOLOGY. Downloaded on September 26,2020 at 11:38:14 UTC from IEEE Xplore. Restrictions apply.

Table 6 - R libraries used in this study

Function	Libraries
Data preparation	Corrplot, dplyr, ggplot2, mosaic and psych
Pre-processing	E1071, Boruta [15]
Machine learning	Caret, class, e1071, gmodels, kernlab, party and randomForest

E. *Evaluation metrics*

The standard measures of accuracy, recall, precision and F-Measure is used to assess the performance of classifiers. Equations 3 to 6 provide the calculation of these measures.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (3)$$

$$Precision = \frac{tp}{tp + fp} \quad (4)$$

$$Recall = \frac{tp}{tp + fn} \quad (5)$$

$$F - Measure = \frac{2tp}{2tp + fp + fn} \quad (6)$$

"Where *tp* is true positive (dropout), *tn* true negative (not dropout), *fp* false positive and *fn* false negative. the researchers consider dropout as the positive class and non-dropout as the negative class. Since the researchers want to minimize false negatives (students who drop out are predicted as students who do not drop out) the researchers will select models with the high recall over those with better precision" [32]. The F-Measure will be analyzed as a trade-off between these metrics.

F. *Evaluation strategy*

For dropout classification, the data set is split in 75% train and 25% test, training the models using grid search and cross-validation on the training set and evaluating it on the test set. All the algorithms are general and enough data is available to conduct experiments for any academic year and any qualification on offer at the University.

IV. RESULTS AND DISCUSSIONS

In this section, the researchers explain the performed experiments for evaluation, which are the best models to predict dropout among the five described classifiers in the previous section. Useful data visualization will be provided in this section for better understanding and interpretation of the results.

Table 7 provides a summary of the balanced dataset that has been split into training and testing data sets based on the classifier. There is equal representation of the classifier in the training and testing data sets. These data sets will be used for the prediction of student dropout through classification.

A. *Dimensional reduction*

The original data set with 19 features (listed in Table 2) including the two-level classifier was reduced to 15 features since the researchers discard to include information in the feature vector such as CALENDER_YEAR, RISK_RATIO, QUALIFICATION and SEMESTER. These mentioned features do not hold predictive power. RISK_RATIO is a derived value from equation 1 to determine the value of FINAL_DECISION. Since the researchers aim to build better, generalized models with the physical properties of the original features, all important features for student dropout prediction were subjected to a relevance test. The feature vector was further reduced to 11 features as listed in Table 8 and indicated in green in Figure 2 with the use of feature selection. The final data set resulted in a feature vector with 12 variables including the FINAL_DECISION.

Table 7- Training and testing data sets

Data set	Not_At_Risk	At_Risk
Reduced dataset *(N=1156)*	578	578
Training dataset *(N=867)*	49.60%	50.40%
Testing dataset *(N=289)*	51.20%	48.80%

Table 8 - Feature selection for dropout

Variable name	Normal Hits	Decision
1. ACCOMMODATION	0.84	Confirmed
2. AGE	1	Confirmed
3. CREDITS_COLLECTED	1	Confirmed
4. DISABILITY	0	Rejected
5. FINANCIAL_AID	1	Confirmed
6. GENDER	0.16	Rejected
7. HOME_LANGUAGE	1	Confirmed
8. MODULES_COMPLETED	1	Confirmed
9. MODULES_FIRST_TIME	1	Confirmed
10.MODULES_REPEATED	1	Confirmed
11.PERSISTENCE	1	Confirmed
12. PREVIOUS_YEAR_ACTIVITY	1	Confirmed
13.YEARS_IN_SYSTEM	1	Confirmed

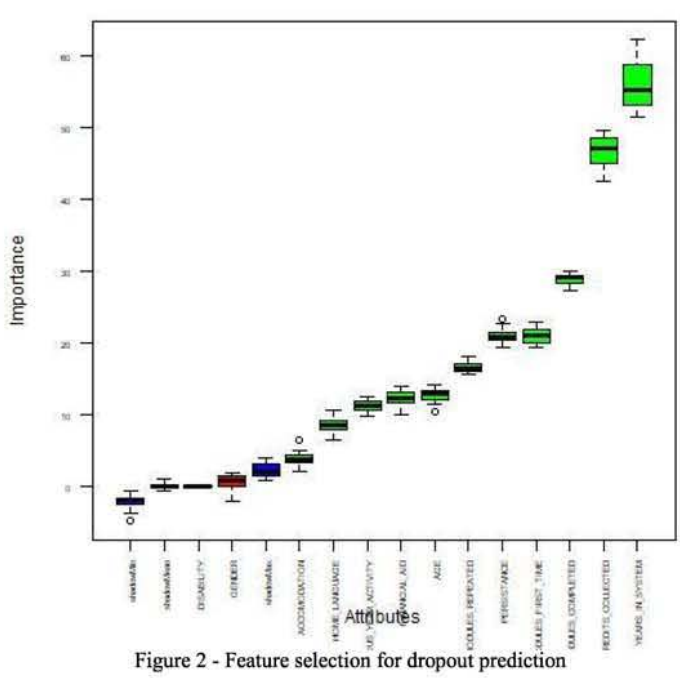


Figure 2 - Feature selection for dropout prediction

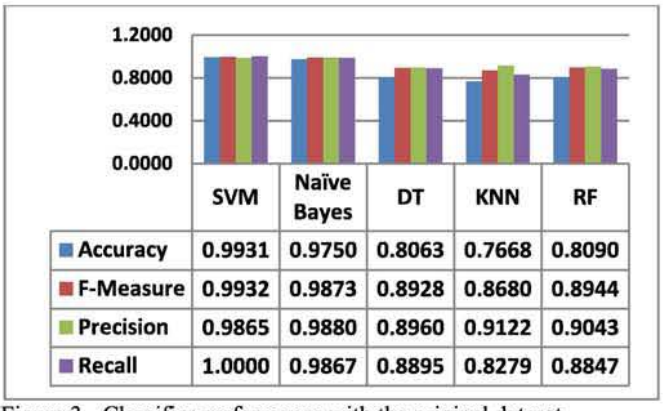
Authorized licensed use limited to: CALIFORNIA INSTITUTE OF TECHNOLOGY. Downloaded on September 26,2020 at 11:38:14 UTC from IEEE Xplore. Restrictions apply.

B. *Classifier performance*

Figure 3 summarizes the evaluation of the five models trained to predict dropout with the original data set corresponding to courses listed in Table 3 respectively.

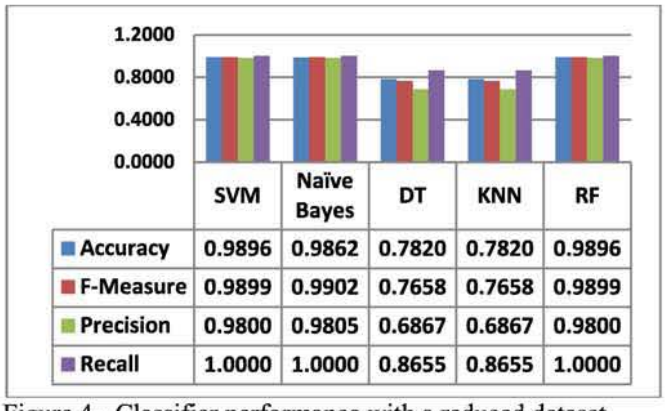
Although all the machine learning algorithms reached an accuracy of more than 75%, SVM and NB performed better when predicting dropout, both obtained an F-Measure score of 99.32% and 98.73% respectively. KNN reached the lowest recall score of 82.79% and RF reached a recall score slightly below DT at 88.47%. The best performing algorithm with the original data set is SVM, which reached the highest score among the classifiers that, was tested with this data set.

Figure 4 summarized the evaluation of the five models trained to predict dropout with a reduced data set through feature selection. KNN, NB and RF had an increase in the accuracy score to 78.20%, 98.62% and 98.96% as the biggest increment. All the other classifiers decreased in accuracy score. SVM and RF performed the same on all measurements in the reduced data set. Only NB measured scores differently to SVM and RF as a pair and DT and KNN as a pair.



	SVM	Naïve Bayes	DT	KNN	RF
Accuracy	0.9931	0.9750	0.8063	0.7668	0.8090
F-Measure	0.9932	0.9873	0.8928	0.8680	0.8944
Precision	0.9865	0.9880	0.8960	0.9122	0.9043
Recall	1.0000	0.9867	0.8895	0.8279	0.8847

Figure 3 - Classifier performance with the original dataset



	SVM	Naïve Bayes	DT	KNN	RF
Accuracy	0.9896	0.9862	0.7820	0.7820	0.9896
F-Measure	0.9899	0.9902	0.7658	0.7658	0.9899
Precision	0.9800	0.9805	0.6867	0.6867	0.9800
Recall	1.0000	1.0000	0.8655	0.8655	1.0000

Figure 4 - Classifier performance with a reduced dataset

V. CONCLUSION

In this paper, the researchers presented a data-driven approach to identify factors for the identification of students at risk of dropout. Through feature selection, factors with no predictive power (noisy data) were eliminated. Different machine learning algorithms were compared and the ones that performed the best were selected. Our classification system's results in dropout prediction are promising, obtaining an F-Measure score of 99.32% and lost some score to 98.99% when the noise in the dataset was removed. The best performing classifier was SVM using the original dataset. This observation was also present when the reduced dataset was applied. A significant improvement has been observed in the RF classifier of the reduced data set.

For future work, the researchers will test our models with new students' data at the University of Technology during the next academic years. In parallel, the researchers will increase the number of students and a variety of degrees to evaluation these models in other scenarios. This would result in a much more complete study.

This study cannot be generalized to all Universities of Technology in South Africa since the analysis is based on sourced data.

ACKNOWLEDGEMENT

The researchers would like to acknowledge the National Research Foundation (NRF, grant number: 105218) for their enabling support of this research paper.

REFERENCES

[1] ADHATRAO, K., GAYKAR, A., DHAWAN, A., JHA, R. & HONRAO, V. 2013. PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS. *International Journal of Data Mining & Knowledge Management Process,* 3(5):39-52.

[2] AULCK, L., VELAGAPUDI, N., BLUMENSTOCK, J. & WEST, J. 2017. Predicting Student Dropout in Higher Education.16-20.

[3] BARADWAJ, B.K. & PAL, S. 2011. Mining Educational Data to Analyze Students' Performance. *International Journal of Advanced Computer Science and Applications,* 2(6):63-69.

[4] BEIKZADEH, M.R., PHON-AMNUAISUK, S. & DELAVARI, N. 2008. Data mining application in higher learning institutions. *International Journal of Informatics in Education,* 7(1):31-54.

[5] BERGIN, S., MOONEY, A., GHENT, J. & QUILLE, K. 2015. Using Machine Learning Techniques to Predict Introductory Programming Performance. *International Journal of Computer Science and Software Engineering (IJCSSE),* 4(12):323-328.

[6] BERGNER, Y. 2017. *Chapter 3: Measurement and its Uses in Learning Analytics.* 1 ed.: Learning Analytics Research Network.

[7] COUNCIL OF HIGHER EDUCATION. 2018. VitalStats Public Higher Education 2016.1 – 124

[8] DIETZ-UHLER, B. & HURN, J. 2013. Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective. *Journal of Interactive Online Learning,* 12(1):17-26.

[9] DEPARTMENT OF HIGHER EDUCATION & TRAINING. Department of Higher Education & Training & Africa, S. 2018. *2000 to 2015 first time entering undergraduate cohort studies for public higher education institutions* [Online]. Available from: http://www.dhet.gov.za/HEMIS/2000%20TO%202015%20FIRST%20TIME%20ENTERING%20UNDERGRADUATE%20COHORT%20STUDIES%20FOR%20PUBLIC%20HEIs.pdf [Accessed: 08 April 2018].

[10] GOVINDARAJAN, K., KUMAR, V., BOULANGER, D. & KINSHUK. 2015, 10-12 Dec. 2015. Learning Analytics Solution for Reducing Learners' Course Failure Rate. *2015 IEEE Seventh International Conference on Technology for Education (T4E).*83-90.

[11] GULATI, H. 2015. Predictive Analytics Using Data Mining Techniques.:713-716.

[12] JAYAPRAKASH, S., MOODY, E., LAURÍA, E., REGAN, J. & BARON, J. 2014. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics,* 1(1), 01/01/:6-47.

[13] JORDAAN, D. & VAN DER MERWE, A. 2015. Best practices for learning analytics initiatives in higher education.53-58.

[14] KABAKCHIEVA, D. 2013. Predicting Student Performance by Using Data Mining Methods for Classification. *CYBERNETICS AND INFORMATION TECHNOLOGIES,* 13(1):61 - 72.

[15] KURSA, M.B., RUDNICKI, W.R. 2010. Feature selection with the Boruta Package. *Journal of Statistical Software,* 36(11):1.

Authorized licensed use limited to: CALIFORNIA INSTITUTE OF TECHNOLOGY. Downloaded on September 26,2020 at 11:38:14 UTC from IEEE Xplore. Restrictions apply.

[16] LATIF, A., CHOUDHARY, A.I. & HAMMAYUN, A.A. 2015. Economic Effects of Student Dropouts: A Comparative Study. *Journal of Global Economics,* 3(2):4.

[17] LEHOHLA, P. 2016. Financial statistics of higher education institutions 2015. (Statistics South Africa), 25 October 2016:36.

[18] LIANG, J., LI, C. & ZHENG, L. 2016, 23-25 Aug. 2016. Machine learning application in MOOCs: Dropout prediction. *2016 11th International Conference on Computer Science & Education (ICCSE).*52-57.

[19] LONTO, M.E., VAN DER WALT, J.S. & CONRADIE, D.P. 2011. PREDICTION OF FIRST YEAR UNIVERSITY STUDENT ACADEMIC PERFORMANCE: AN APPLICATION OF DATA MINING METHODS. In: *2011 International Conference in Data Mining and Knowledge Engineering.*

[20] MAALOUF, M. 2011. Logistic regression in data analysis: An overview. *Int. J. Data Analysis Techniques and Strategies*:1-20.

[21] MATSEBULA, F. & MNKANDLA, E. 2016, 28-29 Nov. 2016. Information systems innovation adoption in higher education: Big data and analytics. *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE).*326-329.

[22] MYTHILI, M. & SHANAV, A. 2014. An analysis of students' performance using classification algorithms. *IOSR Journal of Computer Engineering (IOSR-JCE), 16(1)*: 63-69.

[23] NICOLE, R. "Title of paper with the only first word capitalized," J. Name Stand. Abbrev., in press.

[24] NIKAM, S. 2015. A Comparative study of classification techniques in Data Mining Algorithms. *Oriental Journal of Computer Science & Technology,* 8(1):13-19.

[25] OKEWU, E. & DARAMOLA, O. 2017, 29-31 Oct. 2017. Design of a learning analytics system for academic advising in Nigerian universities. *2017 International Conference on Computing Networking and Informatics (ICCNI).*1-8.

[26] PAPPAS, O.I., GIANNAKOS, M.N. & JACCHERI, L. 2016. Investigating Factors Influencing Students' Intention to Dropout Computer Science Studies. *ITiCSE '16*:6.

[27] PARVIN, H., MIRESMAEIL, M. & HAMID, A. 2015. Proposing a classifier ensemble framework based on classifier selection and decision tree. *Engineering Applications of Artificial Intelligence,* 3719 September 2014:34-42.

[28] PAURA, L., & ARHIPOVA, I. 2014. Cause Analysis of students' dropout rate in higher education study program. *Procedia – Social and Behavioral Sciences*, 109(2014):6.

[29] QIAN, Y. & LEHMAN, J. 2016. Correlates of Success in Introductory Programming: A Study with Middle School Students. *Journal of Education and Learning,* 5(2), March 2016:73-83.

[30] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[31] ROMERO, C., VENTURA, S. & GARCÍA, E. 2008. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education,* 51(1):368-384.

[32] ROVIRA, S., PUERTAS, E., IGUAL, L. 2017. Data-driven system to predict academic grades and dropout. PLoS ONE 12(2): e0171207. doi:10.1371/journal.pone.0171207.

[33] ROY, S. & SINGH, S. 2017, 12-13 Jan. 2017. Emerging trends in applications of big data in educational data mining and learning analytics. *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence.*193-198.

[34] SCLATER, N., PEASGOOD, A. & MULLAN, J. 2016. Learning Analytics in Higher Education, A review of UK and international practice Full report.

[35] SRILEKSHMI, M., SINDHUMOL, S., SHIFFON, C. & KAMAL, B. 2016, 2-4 Dec. 2016. Learning Analytics to Identify Students At-risk in MOOCs. *2016 IEEE Eighth International Conference on Technology for Education (T4E).*194-199.

[36] SRIVASTAVA, M. 2014. Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining. *International Journal of Computer Application (0975-8887),* 88(19):26-29.

[37] STEWARD, S., LIM, D.H. & KIM, J. 2015. Factors Influencing College Persistence for First-Time Students. *Journal of Developmental Education,* 38(3):9.

[38] TANG, J., ALELYANI, S., LIU, H. 2014. Feature selection for classification: A review. *Data classification: Algorithms and applications.*

[39] THOMAS, D. 2013. Factors that Influence College Completion Intention of Undergraduate Students. *Asia-Pacific Edu Res,* 23(2):10.

[40] YUAN, J. & YU, S. 2014. Privacy Preserving Back-Propagation Neural Network Learning Made Practical with Cloud Computing. In: *IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS* [Online]. Available from: [Accessed:

[41] YUKSELTURK, E., OZEKES, S. & TÜREL, Y. 2014. Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. *European Journal of Open, Distance and E-Learning,* 17(1):118-133.

Authorized licensed use limited to: CALIFORNIA INSTITUTE OF TECHNOLOGY. Downloaded on September 26,2020 at 11:38:14 UTC from IEEE Xplore. Restrictions apply.