A Machine Learning Approach to Identifying Students at Risk of Dropout: A Case Study

Roderick Lottering[1], Robert Hans[2], Manoj Lall[3]
Department of Computer Science
Tshwane University of Technology
Gauteng, South Africa

***Abstract*—The increase in students' dropout rate is a huge concern for institutions of higher learning. In this article, classification techniques are applied to determine students "at-risk" of dropping out of their registered qualifications. Being able to identify such students timeously will be beneficial to both the students and the institutions with which they are registered. This study makes use of Random Forest, Support Vector Machines, Decision Trees, Naïve Bayes, K-Nearest Neighbor, and Logistic Regression for classification purposes. The selected algorithms were applied on a dataset of 4419 student records obtained from the institutional database related to Diploma students enrolled in the Faculty of Information, Communication and Technology. The results reveal that the overall accuracy rate of Random Forest (94.14%) was better than the other algorithms in identifying students at risk of dropout.**

***Keywords*—EDM; student dropout; binary classification; ensemble method; KDD**

I. Introduction

Globally, institutions of higher learning have to deal with an increasingly serious problem of student's dropping out of their registered qualifications. Many reasons including absenteeism and financial conditions have been cited for their dropout. The impact of dropout on institutes of higher learning, whether government or privately funded, can be dire as they are often "tuition-dependent". In some countries, including South Africa, the government funding to the institutions is tied to students who graduate. Being able to identify such students, educational institutes can provide a targeted support mechanism to the needy students [1]. Additionally, a high dropout rate is perceived by some as a measure of the quality of educational institutions [2]. To address the problem of student dropout, institutions apply various strategies depending on the perceived student needs and available resources. Examples of strategies put in place to reduce the dropout rate are - assign tutors to needy students, set up learning communities, and provide extended labs access (for practical subjects) [3]. From the discussion presented here, it is apparent that the identification of students at risk of dropout is of significant importance; hence this article aims to formulate a model to address this problem.

Amongst the various approaches adopted to address this problem of identifying students at risk of dropout, educational data mining (EDM) techniques continue to receive great attention. EDM is an area of study to find patterns in educational data through statistics, machine learning (ML), and data mining (DM) algorithms. EDM's aim is to evaluate educational data in order to address the problems of educational research [4]. EDM is interested in the development of methods to evaluate data from educational settings in order to better understand the learners, the learning process and the environment [5] [6].

Data mining and ML continue to receive attention from researchers in diverse fields including education, business and health care. Accordingly, this paper focuses on a comparative analysis of various machine learning techniques including Decision Trees (DT), K-Nearest Neighbor (KNN), Logistic Regression (LR), Naïve Bayes (NB) and Support Vector Machines (SVM) classification algorithms to determine students at risk of dropout. Random Forest as an ensemble method is used to enhance the prediction output of these machine learning techniques.

The institutional guideline in this case study allows a student a maximum of six years to complete the Diploma studies, failing which the student is excluded (forced dropout). Depending on the credits obtained at a certain period during the studies, a student is provided with an opportunity to comply with the academic performance requirements, in order to avoid final exclusion. This study is aimed to determine the students at risk of dropout due to exclusion.

The rest of this article is structured as follows: a literature review is presented in Section II. Section III presents the methodology followed in achieving the research objective. Section IV discusses the results and the conclusion is presented in Section V.

II. Literature Review

According to Yukselturk, Ozekes & Türel [7], data mining has been applied to retrieve data from the various implementations of instructional modes including computer-based, web-based and traditional (face-to-face) education. As indicated by [8], data mining may be used to discover unexpected relationships between student characteristics, teaching strategies and assessments. In the context of online courses, they used Association Rules (AR) to evaluate and produce useful information about dropouts. This model was applied to a Moodle-based Learner Management System (LMS) at the Institute of Computing of a federal university in Brazil with a mix of in-classroom and distant learning students. A total of 27 courses with a student population of 1421 were selected. This population included 242 dropouts.

These study findings showed that reducing the number of dropouts from online courses mediated through the LMS was a relative goal to boost resource utilization where classes for students are small.

Liang, Li & Zheng [2] focused on the student performance in Massive Open Online Courses (MOOCs) using users' behaviour logs. Metadata on classes, course registration records for students and most importantly, user activity logs were collected from their online analytics platform. Their sample of data included thirty-nine (39) courses, with each course containing user activity logs of over 20 000 students over 40 days. Commonly used supervised machine learning algorithms such as LR, NB, SVM, and DT have been used to address this problem. The best performance was the Gradient Boosting Decision Tree (GBDT) with an 88% accuracy.

Adhatrao et al. [9] include merit for examination marks, gender, and marks scored in Science, Technology and Mathematics in the examination of Grade 12 in their dataset to predict student performance. A class label was retained with the expected result, either "Pass" or "Fail". As such, attributes will include distinct values where there was a description of different groups to predict better outcomes. If the merit scored was 120 and above, the merit rating had a "good" value and merit was graded as "bad" if less than 120. This dataset was derived from a university database containing 123 documents.

Aulck et al. [10] analyzed a large, heterogeneous dataset from the University of Washington's Information school. The data included demographic information, school exit information and records from the university. They focused on cohorts over a defined period in a population of 69 116 students. Those who did not complete their studies were marked as dropouts. Three machine learning algorithms (regularized LR, KNN and RF) were applied to the datasets to predict a dropout. The strongest individual predictors of student retention were the Grade Point Average (GPA) in Mathematics, English, Chemistry and Psychology classes. Regularized LR provided the strongest predictions for the dataset.

Bergin et al. [11] reported: "Identifying struggling students at an early stage was not easy as introductory programming modules often have a high student to lecture ratio (100:1 or greater) and early assessment may not be a reliable indicator of overall performance". The factors include: (i) background information, (ii) perceived comfort level factors at the start of the module and (iii) motivation and use of learning strategies. Some of the background factors include among others previous academic experience for example mathematics, science and language. Six different types of algorithms under evaluation included: (i) Logistic Regression, (ii) K-Nearest Neighbor, (iii) Backpropagation, (iv) C4.5, (v) Naïve Bayes and (vi) SVM using Sequential Minimal Optimization (SMO). Three measurement techniques such as overall classifier accuracy, precision and recall were employed in this study. Naïve Bayes produced the highest result among these algorithms in the study.

Whiting et al. [12] included Stochastic Gradient Boosting, RF and rule ensembles (RuleFit) in their approach to implement ensemble methods. Compared to partially adaptive models, the ensemble models provided a better classification and did particularly well. The rule ensemble was appealing in that while providing interpretability, it achieved competitive levels of precision. The dataset was made up of 228 firms, 114 real fraud firms and 114 model companies in the industry.

III. Methodology

In this study, the KDD (see Fig. 1) approach is applied to the dataset in determining the students at risk of dropout. In KDD approach [13], the selected data is subjected to certain preprocessing steps such as removal of outliers and imputation of missing values. Thereafter, dimensionality reduction or transformation techniques to reduce the effective number of variables are performed on the dataset. Subsequently, selected algorithms are applied to the dataset in search of a pattern. The mined pattern is then interpreted to gain knowledge.

A. Dataset

The dataset consisting of 4419 full-time students who enrolled in the Diploma qualifications offered in the Faculty of Information, Communication and Technology (ICT) between 2013 and 2017 academic year was harvested from the institutional database. The normal duration of these Diploma qualifications is three years and requires a student to pass 24 subjects. The total credit value of these 24 subjects is three (3). The dataset consisted of student biographical information and student academic information. Student biographical data includes accommodation indicator, age, disability indicator, financial aid indicator, gender, home language and previous year activity indicator. Student academic record data included qualification, modules and the final mark obtained.

The dataset was enriched with derived values for credits obtained (total number of credits for subjects passed), Number of modules completed, Number of modules repeated (passed on subsequent attempts), Number of modules passed in the first attempt, persistence (a count of modules attempted), years in the system (nr of years in the system from registration) and a final decision class attribute. Accommodation indicates if a student is staying in the university provided residence or not. Persistence is the number of times a student takes a particular subject and therefore measures extra effort a student put into the enrolled studies. An overview of the variables and possible values of the dataset is provided in Table I.

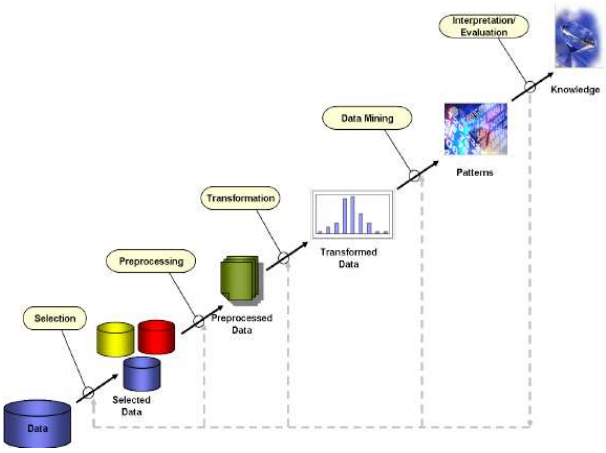


Fig. 1. Knowledge Discovery in Database [14].

TABLE I. OVERVIEW OF DATASET

VARIABLE / FEATURE	VALUES
ACCOMODATION	Y or N
AGE	17 to 48
CALENDER_YEAR	2013 to 2017
CREDITS_COLLECTED	0 to 3480
DISABILITY	Y or N
FINANCIAL_AID	Y or N
GENDER	M or F
HOME_LANGAUGE	AFRI, ENGL, ISIN, ISIX, ISIZ, OTHR, SESO, SESS, SETS, SISW, TSHI, XITS
MODULES_COMPLETED	0 to 55
MODULES_REPEATED	0 to 22
MODULES_FIRST_TIME	0 to24
PERSISTENCE	0 to 70
PREVIOUS_YEAR_ACTIVITY	S or NS
QUALIFICATION CODE	NDIB12, NDIBF1, NDII12, NDIIF1, NDIK12, NDIKF1, NDIL12, NDILF1, NDIP12, DIPF1, NDIS12, NDISF1, NDIT12, NDITF1, NDUI12, NDUIF1, NDIW12, NDIWF1
YEARS_IN_SYSTEM	1 to 7
FINAL_DECISION (CLASS ATTRIBUTE)	PROBATION /EXCLUDED

As per the instructional guideline, a student is allowed a maximum of six years to complete the Diploma or else the student is excluded (forced dropout) from the qualification. The exclusion of poor-performing students is necessary as they impact on the success rate, throughput rate, earnings and reputation of the institution. In order to identify students at risk of exclusion, constant monitoring of the progress of students is essential. Depending on the credits accumulated in a certain time period, a student may continue without any restrictions placed on him or be placed on probation. Probation is essentially a conditional grace period in the exclusion process which provides the student with the opportunity, through specific conditions and interventions, to comply with the academic performance requirements, in order to avoid final exclusion. Table II shows the minimum credit requirements by the students to avoid being excluded or placed on probation. A student that obtains 0.5 credits or more per year is considered to be on the safe side and therefore "NOT AT RISK". A student who obtains less than 0.5 credits is considered "AT RISK" [15].

To derive the values of the class attribute (FINAL_DECISION), Equation 1 was applied to obtain the RISK RATIO.

$$\text{RISK\_RATIO} = 1 - (\text{ACCUMULATED\_CREDITS} / \text{YEARS\_REGISTERED}) \quad (1)$$

If the RISK RATIO is greater than 0.469 and less than 0.55, the value of the class attribute (FINAL-DECISION) is "probation". If the RISK RATIO is greater than 0.55 then FINAL-DECISION is "exclusion". Otherwise, the student is performing satisfactorily and is considered "not at risk" as proposed by Lottering, Hans and Lall [15]. The dataset had instances with the following class categories: “At risk” had 1654 students and the “Not At Risk” category had 2765 records.

B. Preprocessing

All the records with the “Not At Risk” class label was removed before any analysis on the dataset was performed. Preprocessed catered for missing values, outliers and type conversion. The dataset was then subjected to a feature selection process using Regularized Random Forest to reduce the dimensionality of the dataset and consider only the attributes that have some predictive power. Fig. 2 highlights that 9 attributes and their relative importance in the classification process.

TABLE II. CREDIT REQUIREMENTS PER STUDY YEAR

Number of Years registered for a qualification	Maximum Credits obtainable/year	Minimum credits required / year to avoid exclusion.	Accumulated Credits	Risk ratio
1	1.0	0.45	0.45	0.55
2	1.0	0.45	0.90	0.55
3	1.0	0.495	1.395	0.535
4		0.6	1.995	0.50
5		0.66	2.655	0.469
6		0.345	3.00	0.5
Total credits	3.0	3.0		

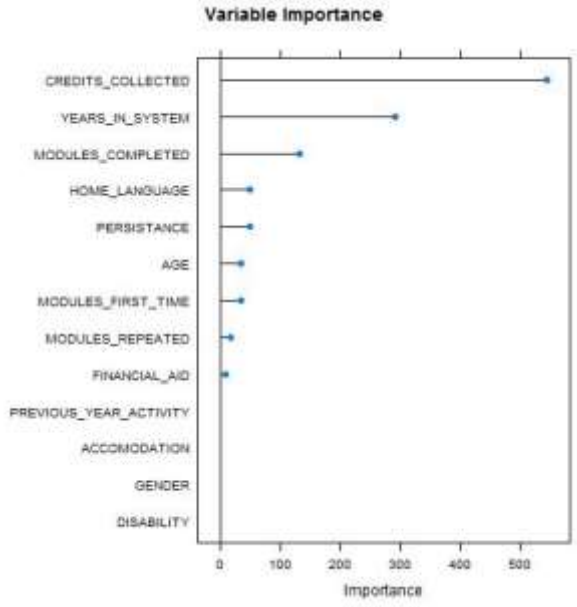


Fig. 2. Feature Selection.

The mean values and standard deviation presented in Fig. 3, contains the range of values reported in Table I. Trends that emerge include a big spread in PERSISTANCE, MODULES_FIRST_TIME and MODULES_COMPLETED due to the standard deviation that is more than 3. The other variables are within the standard deviation of less than 2, indicating a higher concentration around the mean of these features. From the descriptive analysis performed and reported (Table III), it was observed that forty-seven per cent (n=780) had received some form of scholarship. The dominant home language was Sesotho (SESO) which represented 27%. Off these instances, 1074 belonged to the "EXCLUSION" class label compared to the 580 instances that were of the "PROBATION" class label.

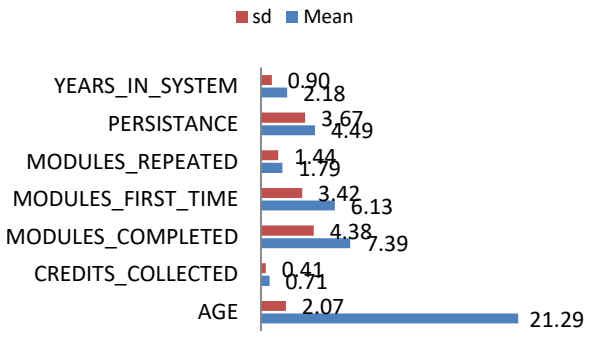


Fig. 3. Mean and Standard Deviation for the Dataset.

TABLE III. DESCRIPTIVE STATISTICS OF THE DATASET

	Frequency	**Percentage**
FINANCIAL_AID		
Yes	780	47%
No	874	53%
HOME_LANGUAGE		
AFRI	3	0%
ENGL	31	2%
ISIN	73	4%
ISIX	72	4%
ISIZ	300	18%
OTHR	96	6%
SESO	440	27%
SESS	126	8%
SETS	148	9%
SISW	120	7%
TSHI	106	6%
XITS	139	8%
FINAL_DECISION		
Probation	580	35%
Exclusion	1074	65%

C. Data Transformation

To classify the students as being at risk of "PROBATION" or "EXCLUSION", the values of attributes listed in Table I, was converted into normalized numerical values. Thereafter, imbalances in the dataset were removed by undersampling the exclusion class label to equal size of the probation class label. Seventy-five per cent (75%) of the dataset was used as a training and validation set while 25% was used for testing purposes. A ten (10) fold cross-validation was used on the training set. Table IV provides an overview of the training and testing balanced datasets for classification purposes.

TABLE IV. TRAINING AND TESTING DATASET

Dataset	Probation	Exclusion
Original balanced dataset (*N=1160*)	580	580
Training and validation dataset (*N=870*)	435 (50%)	435 (50%)
Testing dataset (*N=290*)	145 (50%)	145 (50%)

D. Data Mining

Parmar *et al.* [6] defined machine learning as "computational methods/models using experience (data) to enhance performance". Such programmable computational methods are capable of 'learning' from data and can thus simplify and improve the process of prediction. For the purpose of classification, the following algorithms were used - DT [11], KNN [13], LR [16], NB [11], SVM[11] and RF [8] as ensemble method. These six classifiers are the most suitable classifiers to be used in the identification of students at risk of dropout. A brief explanation of each of these classifiers mentioned above is presented below.

DT are non-parametric classifiers that partitions one feature of a feature vector at a time when the tree's interior nodes correspond to partitioning laws and the class attribute corresponds to the leaf nodes. A vector x function is defined by following the tree starting from the root and applying each node's partitioning rules to decide which branch to follow until a leaf node is reached. The value at the leaf node is classification results [17].

Rovira, Puertas and Iqual [18], define SVM is classification models based on the idea of using hyperplanes to separate data. It considers feature vectors as references in the real Euclidean space. It presupposes that each dot has a single class (0 or 1) and addresses the problem of separating points from any class by constructing the hyperplane from the points of class 0 and class 1 at the greatest distance.

NB is a probability model based on the theorem of Bayes [19]. The algorithm uses Bayes theorem to measure $p(C| x1,.....xn)$ given the n-dimensional function vector of classification class C. In practice, it is assumed that variables are independent.

KNN indicates how many nearby neighbors are supposed to represent the data-point sample class based on the nearest neighbor on k estimates. This kind of learning is 'lazy' as it prevents generalization into the classification stage. The NN algorithm is based on the concept of the probability of the properties of a particular instance in its neighborhood. Each

new instance is compared to existing instances using a distance metric and the new instance is classed according to the majority class of the nearest *K* neighbors [19].

LR uses a logistic function to model a binary dependent variable, although several extensions that are more complex exist. Brownlee [16] explains this as a way to deal with binary classification problems (two-class problems).

RF classifiers are an ensemble learning technique, which creates a set of decision trees, and the performance is the way individual trees are studied. This model is trained with Feature Bagging [17].

E. Evaluation

Accuracy, Kappa, Precision and Recall are the measures that were used to assess classifier performance. Equations 2 to 4 provide definitions of these performance measures.

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (2)$$

$$Precision = \frac{tp}{tp+fp} \quad (3)$$

$$Recall = \frac{tp}{tp+fn} \quad (4)$$

Where *tn* is a true negative, *tp* is true positive, *fn* false negative and *fp* false positive. Dropout is considered the positive class and non-dropout as the negative class. Since we want to eliminate false negatives (students who drop out are expected to be students who do not drop out) we will pick models with the higher specificity over those with better recall.

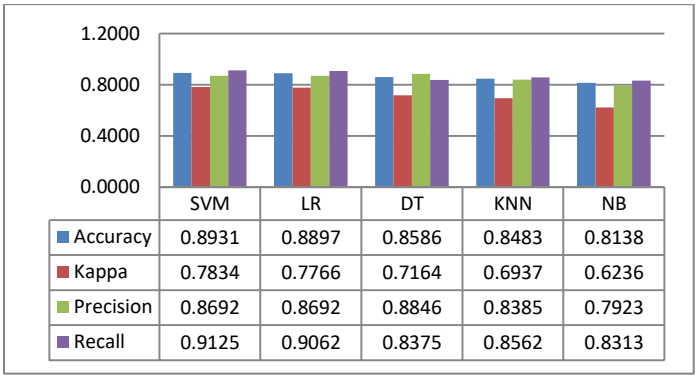
For interrated or intra-rater reliability testing, Kappa is a solid statistics. Its range varies from -1 to +1, where 0 represent a random change and 1 stand for perfect agreement among raters. The result is interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to poor, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as good, and 0.81–1.00 as very good [14].

IV. Results and Discussions

It was observed that NB is the worst performing classifier of the five classifiers under consideration. Fig. 4 presents an overview of the performance of the classifiers. SVM was the best performing classifier among these models. With an accuracy rate of 89.31% and a specificity rate of 91.25%, this classifier is categorised a "substantial" among the raters from a kappa statistic perspective. Recall measured at 86.92%. The best predictor from a precision perspective is Decision Tree, which measured 88.46%. This is the most important measure to select classifiers since the researcher intends to minimise false negatives. KNN was the worst-performing classifier among the supervised machine learning algorithms although it outperformed NB from a precision perspective. The kappa statistic performance of the classifiers is "substantial" among the raters.

The performance indicators of ensemble methods are presented in Fig. 5. Random Forest (RF) had an accuracy of 94.14%, outperforming the initial five classifiers. The RF measured 88.12% from a kappa statistic perspective and attracted an "almost perfect" agreement among the raters.

Fig. 6 presents the performance of the binary classifiers in this study. In general, an AUC of 50% suggests no discrimination (ability to predict students at risk of probation or exclusion based on the test). All the classifiers are therefore categories as "outstanding" since all of them have a measure of more than 90%. The AUC score presents an aggregate measure of performance across all classification thresholds. Random Forest is the best performing classifier with a 99% measure and DT and KNN measured 91% as the least favourable classifier.



	SVM	LR	DT	KNN	NB
Accuracy	0.8931	0.8897	0.8586	0.8483	0.8138
Kappa	0.7834	0.7766	0.7164	0.6937	0.6236
Precision	0.8692	0.8692	0.8846	0.8385	0.7923
Recall	0.9125	0.9062	0.8375	0.8562	0.8313

Fig. 4. Overall Classifier Performance.

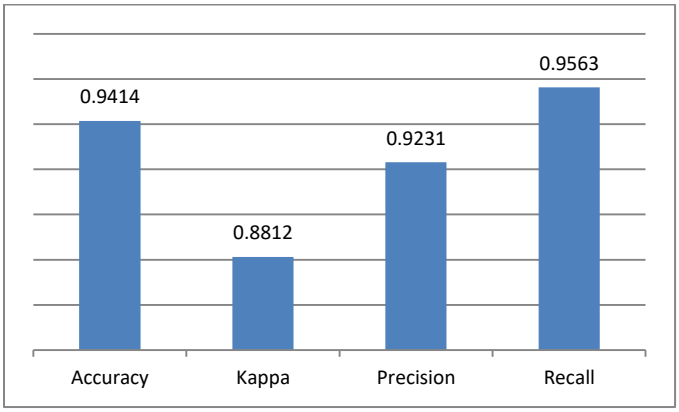


Fig. 5. Performance of Random Forest.

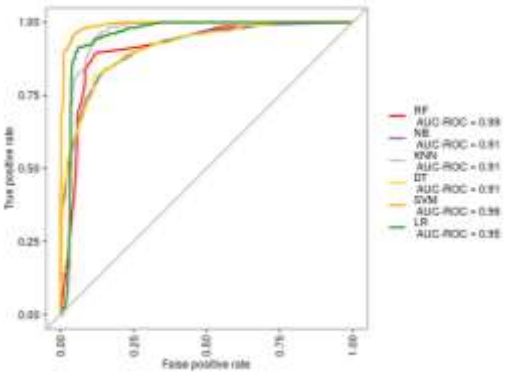


Fig. 6. ROC Curves for Classifiers with a Reduced Dataset.

V. CONCLUSION

In this paper, we examined the factors that could be used to identify a student at risk of dropout at a university of technology. The Tshwane University of Technology was used at a case study. Data of fulltime students for the academic years 2013 to 2017 in various course offerings in the Faculty of Information and Communication Technology was harvested from the institutional database. The data mining process was accomplished by applying the KDD approach. We applied Decision Trees, K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Support Vector Machines and Random Forest for classification purposes. It was observed that Random Forest outperformed the other classifiers for the given dataset. The accuracy achieved by the Random Forest model was 94.14%.

For future work, these models will be tested with new students' data over a longer period. In parallel, the number of students and a variety of degrees will be increased to evaluate these models in other scenarios. Although this research can predict students at risk of dropout, the conclusion cannot be generalized as the data is from a specific University of Technology.

ACKNOWLEDGMENTS

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

REFERENCE

[1] F. MAKOMBE AND M. LALL. A predictive model for the determination of academic performance in private higher education institutions. International Journal of Advanced Computer Science and Applications (IJACSA). 11(9). 2020.

[2] LIANG, C. LI, and L. ZHENG, Machine learning application in MOOCs: Dropout prediction, Paper presented at the 2016 11th International Conference on Computer Science & Education (ICCSE) 23-25 Aug 2016.

[3] B.O. BAREFOOT, Higher education's revolving door: confronting the problem of student drop-out in US colleges and universities, Open Learning: The Journal of Open, Distance and e-Learning, 19:1, 9-18, DOI: 10.1080/0268051042000177818, 2004.

[4] T. BARNES, M. DESMARAIS, C. ROMERO and S. VENTURA, Presented at the 2nd Int. Conf. Educ. Data Mining, Cordoba, Spain. 2009.

[5] R. BAKER, "Data mining for education," in International Encyclopedia of Education, B. McGaw, P. Peterson, and E. Baker, Eds., 3rd ed. Oxford, U.K.: Elsevier, 2010.

[6] C. PARMAR, P. GROSSMANN, J. BUSSINK, P. LAMBIN and H.J.W.L. AERTS, Machine Learning methods for Quantitative Radiomic Biomarkers, Scientific Reports, 5:13087, 2015.

[7] E. YUKSELTURK, S. OZEKES and Y TÜREL, 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, 2014.

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

[9] K. ADHATRAO, A. GAYKAR, A. DHAWAN, R. JHA and V. HONRAO, Predicting students' performance using id3 and c4.5 classification algorithms. International Journal of Data Mining & Knowledge Management Process, 3(5):39-52. 2013.

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

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

[12] D.G. WHITING, J.V. HANSEN, J.B., MCDONALD, C. ALBRECHT and W.S. ALBRECHT, Machine learning methods for detecting patterns of management fraud. Computational Intelligence, 28(4):24, 2012.

[13] S.A. ALASADI and W.S. BHAYA, Review of Data Processing Techniques in Data Mining. Journal of Engineering and Applied Sciences, 12(16): 4192 – 4107, 2017.

[14] A. GUERRA-HERNÁNDEZ, R. MONDRAGÓN-BECERRA and N. CRUZ-RAMIREZ, Explorations of the BDI multi-agent support for the knowledge discovery in databases process. Research in Computing Science 39, 221-238, 2008.

[15] R. LOTTERING, R. HANS AND M LALL. A model for the identification of students at risk of dropout at a university of technology. 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD). IEEE, 2020.

[16] J. BROWNLEE, Logistic regression for Machine learning. Machine Learning Algorithms, https://machinelearningmastery.com/logistic-regression-for-machine-learning, 2016, Accessed 19/06/2020.

[17] R. BOST, A. RALUCA, T. STEPHEN and G. SHAFI, Machine Learning Classification over Encrypted Data [Electronic Version], 2015.

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

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