A Machine Learning Approach for Predicting Students' Second Year Outcomes

Shiluva Claudia Kubayi, Ashwini Jadhav, and Ritesh Ajoodha

Faculty of Science
University of the Witwatersrand
Johnnesburg South Africa
{shiluva.kubayi1@students.wits.ac.za,
ashwini.jadhav@wits.ac.za,
ritesh.ajoodha@wits.ac.za}

Abstract. The number of students enrolling in higher education institutions in South Africa has increased from the year 1994 , however many of these students get academically excluded . Machine learning techniques , along with statistical analysis and data mining are one of the most important ways to study student performance and success . This paper aims at forecasting students second-year outcomes, to deduce if they are at risk of getting academically excluded or will proceed to register for the following academic year . This way , students who are at risk can be provided with support to avoid being academically excluded . Six predictive models , namely , the K-Nearest Neighbors, Random forest, Decision trees , Naive Bayes , Logistic regression and Multi-layer perceptron were trained . The Random forest proved to be a good classification model amongst the others with an accuracy of 83 % , precision of 83 % , recall of 82 % and an F1 score of 83 % . The significance of this study is to promote student success initiatives in higher learning institutions to enhance throughput rates .

Keywords: academic exclusion, at risk, machine learning, predicting student outcomes

1 Introduction

Acceptance into a university program is frequently a life-changing opportunity for most South African matriculants . This is because it assures a greater income and a better quality of life. Regrettably, the majority of students who are accepted into university programs do not complete their degrees . Relatively, nearly 25% of students are forced to withdraw from higher education institutions each year in South Africa for the reason that they are excluded on academic grounds [13] . Academic exclusion occurs when one is forced to leave university because they did not perform well academically , i.e. , one did not pass enough courses to meet the minimum credit criteria to register for the next year . To hold against such issues, student protests and

boycotts usually takes place . These protests sometimes lead to lives being lost and also disturb the academic year .

A study by [7] showed that students from the faculty of Science , and faculty of Engineering get academically excluded from their second year of study , while those in the faculty of Accounting get academically excluded from their third year of study . A study by [15] studied the factors influencing graduation in engineering students , while a study by [10] aimed it's research at learning more about the institutional elements that influence student retention from their first to second year . Vulnerable students are defined in several studies including [12] as those who drop out from a university program or fail out . Vulnerable students will be defined in this study as those who are at risk of getting academically excluded .

The goal of this research is to make stakeholders aware of vulnerable students in the early stages of their academic journey . These stakeholders includes university management , lecturers , students funders , Student Representative Councils and students themselves . The aim of this paper is to forecast students second-year outcomes , we will look at two outcomes, proceed and excluded . Through the predictive models , stakeholders will be able to play part in decreasing the number of vulnerable students through student success initiatives that can be employed on vulnerable students .

This paper uses data that was synthetically generated by a Bayesian network to train six predictive models to predict the students' second year outcomes . When training the data , the attributes were chosen based on their usefulness in achieving our goals and objectives . Predictive model evaluation was done , using the accuracy , f1 score , confusion score , recall and precision .

The six predictive models trained in this study are the K-Nearest Neighbors , Random forest , Decision trees , Naive Bayes , Logistic regression and Multi-layer perceptron . After training our six predictive models and applying evaluation metrics , the results showed that Random forest was the best when it comes to predicting students outcome with an accuracy of 83 % , recall of 82 % , precision of 82 % and F1 score of 83 % . The K-Nearest Neighbours was the least performing model with an accuracy of 73 % , recall of 73 % , precision of 73 % and F1 score of 74 % . The random forest accurately predicted the outcome 10 % more than the K-Nearest Neighbours .

The contribution of this research is to present a predictive model to forecast students' second-year outcomes . Higher education institutions will have an early understanding of which students are at risk of academic exclusion . This will make it possible to implement and promote early interventions that will help with students success .

The following sections make up this paper: section 2 examines prior research , followed by section 3 which examines the methodology utilized to perform the applicable experiments , and section 4 gives the qualitative results of the study as well as analysis of the results . Finally , section 4 presents a summary of the paper as well as it's significance .

2 Related Work

This section comprises of three subsections . The first inquire into the most influential models of student performance , the second inquire into features which have been shown to influence student performance as read in literature , and the last focuses on the evaluation metrics employed in prior work .

2.1 Attributes affecting students performance

Student performance is a vital or rather crucial metric employed to track students and institutional goals [4] . In this paper , we employ the conceptual framework model by [16] , here the author is concerned with the background , pre-college and individual attributes to deduce students performance . Studies by [1] , [4] , [5] have deduced that biographical attributes have a huge impact on students performance followed by individual attributes , on the other hand , [9] and [8] deduced that personal attributes have a higher impact on students followed by pre-college attributes .

2.2 Predictive Models

The adaptivity of machine learning models , as well as their ability to integrate large and complex dataset , has led to the adoption of machine learning over traditional statistical models in a number of studies . A significant number of predictive models have been employed in the past to determine students performance . The random forest is the most utilized predictive model, this model can be found in the paper by [2] , [3] , [1] , [4] and [5] . It was found to have a high accuracy in the paper by [2] , [1] and [4] .

Studies by [9] and [8] have deduced the Decision tree to be the most accurate model. Other models that have been employed in prior work include C4.5 Decision tree, as done by [9] and [4] and , Multi layer perceptron as done by [3] , [1] and [5] . The list of models employed in the literature is exhaustive , this study will be using the following model: K-Nearest Neighbors , Random forest , Decision trees , Naive Bayes , Logistic regression and Multi-layer perception , as supported by prior work. These models will be defined and explained in the methodology section .

2.3 Evaluation Metrics

Several studies have employed the accuracy as an evaluation metric [1] , [4] and [5] . Accuracy is defined as the ratio of correct prediction to evaluate how many valid guesses there are , dependent on the the overall amount of prediction . In the field of student performance predictions , the Area Under the Curve (AUC) and Receiver Operating Characteristic (ROC) curves are also one of the critical metrics . [4] is among the many authors that have employed this metric in their model evaluation , including the Kappa statistics .

The value of the data supplied by a predictive model concerning the predicted and actual class is stored in the confusion matrix . The following authors employed the confusion matrix as an evaluation method , [2] , [3] , [4] and [5] . More about the confusion will be discussed in the methodology section .

The F1 score is crucial for examining the efficiency of the classifiers . Authors including [2] and [3] have employed the F1 score together with the recall and precision . This paper will employ the accuracy , confusion matrix , f1 score , recall and precision as the evaluation metric . More about the evaluation metrics will be learned in the next section .

3 Methodology

This paper employs synthetic data set to train six predictive models . These predictive models are K-Nearest Neighbors , Random forest , Decision trees , Naive Bayes , Logistic regression and Multi-layer perceptron . We want to pindown which predictive model best deduces the outcomes , 'Excluded' and 'Proceed' .

3.1 Data Collection and Pre-Processing

This study utilizes data that was synthetically generated by a Bayesian network . A Bayesian network is defined as a kind of probabilistic model that utilises a directed acyclic graph to illustrate a class of variables and how they correlate . Each node in the graph illustrates a feature while each edge uniting them illustrates the correlation between them . Nodes that are not united illustrates features that are conditionally independent of each other . The node that the edge emerges from is called the parent node whereas , the node being directed to is called the child node, with the edge forming a conditional causal link to the child node from the parent [3] . The casual relationship makes use of the Bayes Theorem , which is described in respect of conditional probability :

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \tag{1}$$

In equation 1 above, A, together with B are instances

P(A | B) is defined as the probability of having instance A given that instance B holds , P(B | A) is defined as the probability of having instance B given that instance A holds , P(A) and P(B) are the independent probabilities of both A and B respectively.

Figure 1 illustrates the Bayesian network structure that was utilised in generating the data used in this research. When generating the data , the forward sampling algorithm was used . The values of the parent nodes are sampled from a conditional distribution , and the values of the children nodes are sampled from their corresponding parent-sets . The sampling procedure is topologically ordered and iterative until all node values are generated . Because

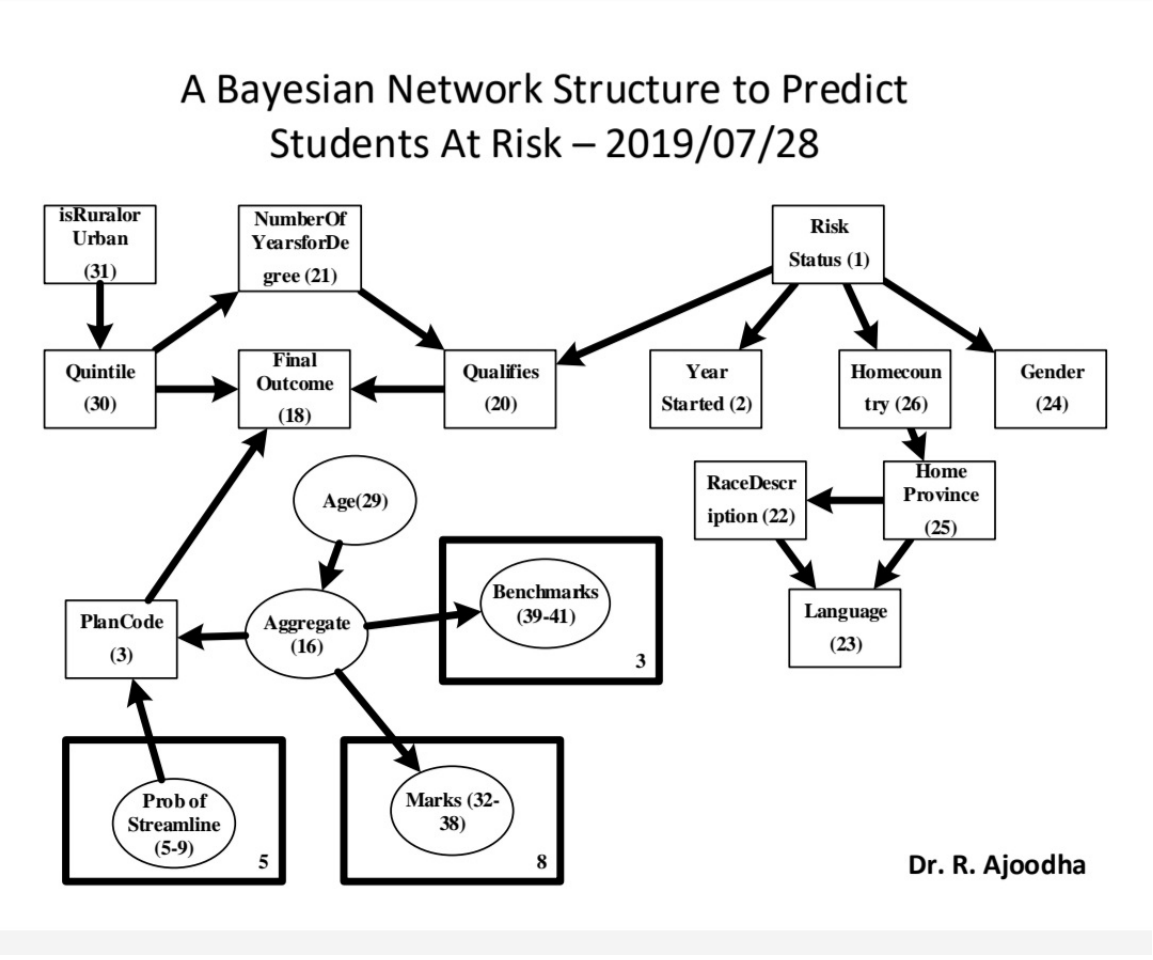


Fig. 1. Network used for generating data

synthetic data was employed , no ethical considerations were required . The limitation encountered when implementing the research methodology is that the data had a lot of missing data .

The data set was presented in the form of an Excel spreadsheet . During data pre-processing , nonsensical data was drawn out, this includes marks that were above 100 % . Columns that were not relevant to our study were also removed from the spreadsheet , this includes third-year outcomes , age at third year , etc . A total of 41 qualities were gathered , with approximately 50 000 sampled observations . Variables were reduced to 11 and observation to 1945 after data cleaning and feature selection . Table 1 shows a condensation of these variables . The Excel spreadsheet was imported to Juypter notebook, which uses Python by the use of Pandas . Data sets with numerous missing values were dropped and the Labelencoder function was used to encode variables that were 'objects' , this includes gender , home province , rural or urban , plan code , race and first year outcome . The data was divided into two parts : 70 % was used to train our models , and 30 % was used to test the models . The sklearn metrics was used to implemnt the algorithms . The success of an academic year was measured using the variable Outcome , and it could take on either Proceeded or Excluded .

3.2 Features

This research uses the conceptual framework found in [16] . The framework assumes that the factors that influence a student's academic performance are as follows :

1. Biographical characteristics
2. Pre-College observations
3. University Enrolment observations

The features were chosen based on their use in achieving our goals and objectives . Table 1 gives the features considered under each category.

Table 1: Features used in this study

Biographical Characteristics	**Pre-College Observations**	**University Enrolment Observations**
Rural or Urban	NBTAL	Agrregate YOS1
Home Province	NBTMA	First year outcome
Age at first year	NBTQL	Plan Code
Race		
Gender		

Table 2 below gives the description of the features and their possible values .

Table 2: Description of the features used

Features	Description	Possible Value
Rural or Urban	The school location	{rural, urban}
Home province	The province the student originates	{All 9 South African provinces, and other national states }
Age at first year	Age of the student at their first year of study	{15 to 60}
Race	Racial description	{Black, White, Indian, Coloured, Chinese}
Gender	Sex identity	{Female, Male}
NBTAL, NBTMA, NBTQL	National Benchmark Tests	{0 to 100}
Plan code	Code for career choice	{All science career choices code}

3.3 Classification Models

The following predictive models were trained by the data .

- **K- Nearest Neighbors:** A concise algorithm for identifying a data point constructed on the classifications of the K neighbouring points around it , supposing that data points with similar characteristics are clustered together . The efficiency of this technique is determined by the K value selected in addition to the distance metric , and because it is particularly delicate to outliers , several values of K are assessed to obtain the optimum predicted results [4] .
- **Random Forests:** Commonly known as random decision forests . They are a classification and regression ensemble learning method . They works by fitting numerous decision tree classifiers on different sub-samples of the dataset and , employ averaging to increase forecast accuracy and restrict data over-fitting . The random forests executed on this research follows the one executed in RF .
- **Decision Trees:** A type of non-cyclic flowchart . They are one of the widely used approach when working with inductive inference . They are made up of internal nodes , which correlates to a logical test on feature , and the coupled branches that portray an outcome of the test . The branches and nodes that constructs this model are a kind of non-cyclic flowchart DT .
- **Naive Bayes:** For most classification problems , Naive Bayes are the most dynamic and rational learning algorithms . They are established on Bayes' idea of compelling assumptions of independence within features , which is implemented in a Bayesian framework NB .

- **Logistic Regression:** It is one of the most common and straightforward models for modelling the linear association linking a dependent variable (Y) and independent variables (X_i) , where i denotes a feature . The term 'logistic' have relevance to a categorical response variable that is binary or dichotomous for two categories (binomial/binary logistic regression) . Multi-nomial logistic regression , on the other hand , could have more than two classes [4] .
- **Multi-layer Perceptron:** It is a deep learning model with numerous layers of input nodes that are coupled as a directed graph connecting the input and output layers . It is a feedforward artificial neural network that employs backpropagation to train and generate a collection of outputs from a collection of inputs [17] .

3.4 Model Evaluation

The employment of evaluation is to assess the functionality of the six models in order to figure out the model with the best results . K-fold cross-validation was utilised in this investigation , using K=10 , K being the number of groups to divide the data into . The data is then randomly divided into K equal-sized sections . We have the following,

- **Recall** is described as the amount of the total number of excellent instances that were actually found . It is also referred to as the true positive rate . It is given by the equation below

$$Recall = \frac{TP}{FP + TP} \tag{2}$$

- **Precision** is described as the ratio of relevant occasions amid the retrieved occasions . It is given by

$$Precision = \frac{TP}{FN + TP} \tag{3}$$

- **F1 score** is described as the weighted harmonic mean of a test's precision and recall . It presents a more pragmatic option for evaluating the effectiveness of a test than recall or precision independently as it stabilizes the usage of both . It evaluates the efficiency of a model . It is given by

$$F1 - score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4}$$

- **Accuracy** is determined as the quantity of correct forecasts based on the overall amount of forecasts . It is given mathematically as

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

– **Confusion Matrix** is applied to evaluate the favorable outcomes of machine learning algorithms . It displays which ratio of a binary variable was accurately and inaccurately classified . Below is an example of a confusion matrix.

		Actual Class	
		Positive (**P**)	Negative (**N**)
Predicted Class	Positive (**P**)	True Positive (**TP**)	False Positive (**FP**)
	Negative (**N**)	False Negative (**FN**)	True Negative (**TN**)

4 Results and Analysis

In this section , we will go over the findings from the six predictive models used in this paper in deducing 'proceed' and 'excluded' . Random forest outperformed all the models with an accuracy of 83 % .

4.1 Classification Model

The findings of the six classification models are explained in this sub-section . The metrics that were employed in this paper to assess the accurateness of the classification models are classification accuracy , confusion matrices , recall , precision and f1 score . Each of them is discussed below :

Classification Accuracy: We used accuracy to evaluate how many of the forecasts were correctly categorized out of all the ones made . Table 3 gives a summary of each of the models employed in this paper and how they are accurately classified .

Table 3: Accuracy of the predictive model Summary

Predictive Models	**Accuracy(%)**
Random Forest	83
Naive Bayes	77
Decision Trees	76
Multilayer Perceptron	79
Logistic Regression	76
K-Nearest Neighbours	73

Confusion Matrices This evaluation metric is used to quantify the degree of ambiguity between the classification classes . It is an N by N matrix where the

columns are anticipated class labels and the rows are actual class labels . Tables 4 to 9 shows the confusion matrix of each of the models .

		Predicted	
		Excluded	Proceed
Actual	Excluded	676	212
	Proceed	128	929

Table 4: Confusion Matrix for Random Forest

		Predicted	
		Excluded	Proceed
Actual	Excluded	674	214
	Proceed	226	831

Table 5: Confusion Matrix for Naïve Bayes

		Predicted	
		Excluded	Proceed
Actual	Excluded	683	205
	Proceed	266	791

Table 6: Confusion Matrix for Decision Trees

		Predicted	
		Excluded	Proceed
Actual	Excluded	676	212
	Proceed	128	929

Table 7: Confusion Matrix for Multilayer Perceptron

		Predicted	
		Excluded	Proceed
Actual	Excluded	620	268
	Proceed	199	858

Table 8: Confusion Matrix for Logistic Regression

		Predicted	
		Excluded	Proceed
Actual	Excluded	623	265
	Proceed	253	804

Table 9: Confusion Matrix for K-Nearest Neighbours

We also obtained that the K-Nearest Neighbour has the highest misclassificationn rate of 27% , followed by Decision tree and Logistic regression with 24%, Naive Bayes with 23%, Multilayer perceptron coming second last with 21% and lastly, random forest with 17% .

Precision the greater the value of the precision , the proficient the model is in forecasting admissible outcomes . Table 10 contains the results of the precision metric for the classification models .

Table 10: Precision score of the models

Predictive Models	**Excluded(%)**	**Proceed(%)**
Random Forest	84	81
Naive Bayes	75	80
Decision Trees	72	79
Multilayer Perceptron	82	78
Logistic Regression	76	76
K-Nearest Neighbours	71	75

Overall , the random forest has the best precision of 0.83 , followed by Multilayer perceptron with 0.8 and then the rest of the models follows in this order, Naive Bayes (0.78) , Decision trees (0.76) , Logistic regression (0.76) and K-nearest neighbours (0.73) .

Recall this is also referred to as sensitivity . Table 11 illustrates the results of the recall metric for classification models and different outcomes . The recall is beneficial to us in understanding which classification models accurately classify the outcomes and which outcome have a higher percentage of accurately classified .

Table 11: Recall score of the models

Predictive Models	**Excluded(%)**	**Proceed(%)**
Random Forest	76	88
Naive Bayes	76	79
Decision Trees	77	75
Multilayer Perceptron	70	87
Logistic Regression	70	81
K-Nearest Neighbours	70	76

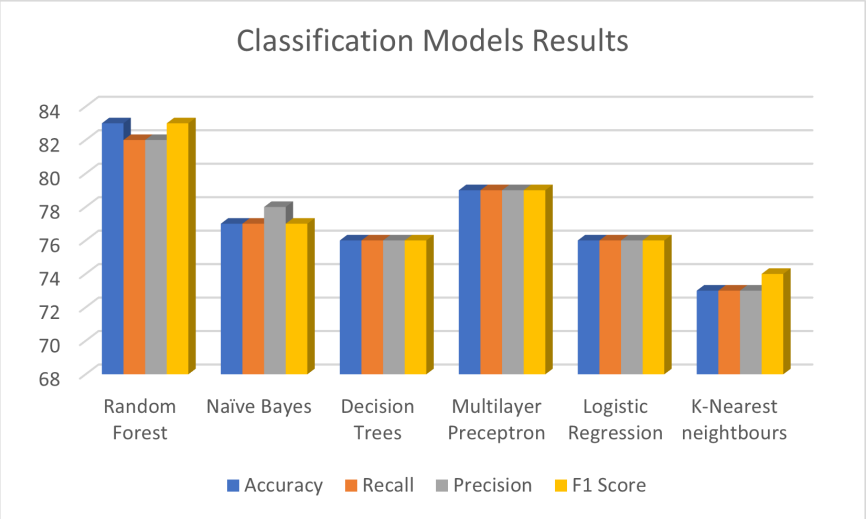
F1 Score Table 12 gives a summary of the F1 score for the models and the outcomes.

Table 12: F1 score of the models

Predictive Models	**Excluded(%)**	**Proceed(%)**
Random Forest	80	85
Naive Bayes	75	79
Decision Trees	74	77
Multilayer Perceptron	76	82
Logistic Regression	73	79
K-Nearest Neighbours	71	76

Compared to the other classifications , Random forest has the highest f1 score at 0.83 , followed by Multilayer perceptron with 0.79 . The rest of the

models follows in the following order , Naive Bayes (0.77) , Decision tree (0.76) , Logistic regression (0.76) and lastly K- nearest neighbours (0.74) . From this , we learn that the K-Nearest Neighbour is not a good model in classifying the outcomes . The graph below illustrates a summary of all the results obtained from the classifications models and the overall evaluation metrics results .



4.2 Analysis

The limitations we encountered during this work was that synthetic data may not express real life systems and observations , which distort the results . However, it does allow us to create a theoretical scenario in which a proof of concept can be created . This raises concerns about the veracity of the data and the applicability of such models in the real world .

We found that the Random forest better predicted the students' outcomes compared to the other five classification models , with an accuracy of 83% . The K-nearest neighbours had the lowest accuracy with 73% . The greater performance of the Random forest compared to the rest , is due to the fact that the model is a proficient model when there are numerous missing values , which was the case with our data . This model establishes multiple different trees in which each of them constructs their own forecasts and the class that is the mode of the individual trees turns out to be the predictive model employed in testing . The Multilayer perceptron preformed second best , this may be due to it using backpropagation to train and generate a collection of outputs derived from a collection of inputs .

K-Nearest Neighbours having a low accuracy compared to the other models is due to it being sensitive to outliers , and also due to the assumptions it makes . It assumes that similar data points are near each other , which was not the case in our data . We used different values of K to improve the model , but it was still the least performing .

5 Conclusion and Discussion

We aimed our study at predicting students second-year outcomes as to deduce if they will be academically excluded or they will proceed to register for the following academic year . The outcomes to be predicted were 'excluded' or 'proceed' . This is after a study by [7] showed that students get academically excluded in their second year of study in the faculty of science and Engineering . We have seen the negative impact student exclusion have on stakeholders and society in general in the past . We used synthetically generated data to train six predictive models , and we found that the Random forest showed better results compared to Decision Trees , K-Nearest Neighbours , Naive Bayes , Multilayer perceptron and Logistic regression .

The importance of this paper is to early detect students who are at risk of getting academically excluded in their second year of study, by providing a predictive model stakeholders can use to provide intervention to students who need support . Future work can include using real data on South African high institution students , that way we can look at different faculties , and the features that impact student academic exclusion in respective faculties . The results of this paper when applied to real-life data could come in handy to higher education institutions in helping decrease the academic exclusion rate .

Acknowledgement

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant number: 121835).

References

1. Buraimoh, E,, Ajoodha R., Padayachee K.: Prediction of Student Success using Student Engagement with Learning Management System. In: Interdisciplinary Research in Technology and Management, 577–583. CRC Press, (2021).
2. Mngadi, N.: A theoretical model to predict undergraduate learner attrition using background, individual, and schooling attributes. PhD diss., (2020)
3. Philippou, N., Ajoodha R., Jadhav A.: Using machine learning techniques and matric grades to predict the success of first year university students. In: 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC), pp. 1–5. IEEE, 2020.
4. Mngadi, N., Ajoodha, R., Jadhav, A.: A Conceptual Model to Identify Vulnerable Undergraduate Learners at Higher-Education Institutions. In: 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC) 1–8. IEEE (2020)
5. Abed, T., Ajoodha, R., Jadhav, A.: A prediction model to improve student placement at a south african higher education institution. In: 2020 International SAUPEC/RobMech/PRASA Conference, pp. 1–6. IEEE, (2020)
6. Asif, R., Agathe M., Syed A.A., Najmi G.H.: Analyzing undergraduate students' performance using educational data mining. Computers and Education 113, 177-194 (2017)

7. Rooney, C., Van Walbeek, C.: Some determinants of Academic Exclusion and Graduation in three faculties at UCT. (2015)
8. Yehuala, M. A.: Application of data mining techniques for student success and failure prediction (The Case Of Debre Markos University). International journal of scientific technology research, 4(4), 91–94 (2015).
9. Pal, A. K., Pal, S.: Analysis and mining of educational data for predicting the performance of students. International Journal of Electronics Communication and Computer Engineering, 4(5), 1560–1565 (2013).
10. Poggendorf, B. P.: Exploring predicted vs. actual first-to-second year retention rates: A study of evangelical Lutheran church in America colleges. PhD diss., (2013)
11. Pham, D.T., Ruz G.A.: Unsupervised training of Bayesian networks for data clustering. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2109, 2927–2948 (2009)
12. Anand, P., Herrington, A., Agostinho, S.: Constructivist-based learning using location-aware mobile technology: an exploratory study. In: EdMedia+ Innovate Learning . Association for the Advancement of Computing in Education (AACE). 2312–2316 (2008)
13. Koen, C., Cele, M., Libhaber A.: Student activism and student exclusions in South Africa. International Journal of Educational Development. 26, 404–414 (2006)
14. Pal, M.: Random forest classifier for remote sensing classification. International journal of remote sensing 26, 217-222 (2005)
15. Zhang, G., Anderson, T. J., Ohland, M. W., Thorndyke, B,R.: Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. Journal of Engineering education, 93(4), 313–320(2004)
16. Tinto, V.: Dropout from higher education: A theoretical synthesis of recent research. Review of educational research. 45, 89–125 (1975)
17. Waikato Enviroment for Knowledge Analysis, https://weka.sourceforge.io/doc.dev/weka/classifiers/functions/MultilayerPerceptron.html