Educational Data-mining to Determine Student Success at Higher Education Institutions

Ndiatenda Ndou
School of Computer Science and Applied Mathematics
The University of the Witwatersrand,
Johannesburg, South Africa
ndiatenda.ndou@students.wits.ac.za

Ritesh Ajoodha
School of Computer Science and Applied Mathematics
The University of the Witwatersrand,
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

Ashwini Jadhav
Faculty of Science
The University of the Witwatersrand,
Johannesburg, South Africa
ashwini.jadhav@wits.ac.za

***Abstract*—The expansion of enrolments in South African higher education institutions has not been accompanied by a proportional increase in the percentage of students who graduate. This is an ongoing problem faced by the Department of Higher Education and Training in South Africa (DHET). In their 2020 undergraduate cohort studies, DHET reported that the percentage of first time entering students graduating in minimum allocated time from 3 year degrees has remained low, ranging between 25.7% and 32.2%, for the academic years 2000 to 2017. This indicates students are struggling in higher education, as more than 60% of students being admitted by the system are consistently not completing their chosen field of study in the allotted time. In this study, we introduce an approach that involves prediction of student performance at each year of study until qualifying, for students at a South African higher education institution. The present study applies various classification techniques to a synthetic data-set, generated by a Bayesian network, with the aim to show that these classifiers can be used to predict student performance in advance with the aim to promote student success and avoid the negative consequences of students struggling to complete their studies or dropping-out altogether.**

***Index Terms*—Student Performance, Prediction, Higher-Education, Machine Learning, Socioeconomic, Psycho-Social.**

I. Introduction

Time at the university is of significant value and produces a return for the student who completes their degree. However, when further research is done on the influence of higher education on the lives of young adults entering the system, a more complex but interesting picture emerges. A study conducted in 2018 revealed that time spent at university is not in vain for those students who have not yet completed, or may not complete their chosen field of study [8]. Although the students have not achieved the goal they set to achieve upon registration, they derive varied and complex benefits from the university environment which expands the way they see the world, the roles they can imagine themselves playing in the world, and in many instances, it changes how they view themselves and other individuals, as well.

The benefits that can come with the higher education system extend far beyond the degree holder. A study was done on the association between human capital, universities, and quality of life found that local level of human capital (measured by the percentage of adults with a degree), and higher education institutions, bring to life valuable consumption amenities that increase the quality of life in areas, where the quality of life is determined by differences in real wages [32]. These respective findings show that what goes on in higher learning institutions has broader relevance and is of societal importance, thus not limited to just individual or economic advancement, and therefore, pointing out the clear need to explore and understand the activities in the institutions [8].

Although the higher education system is a system of great rewards to individuals, society, and the economy, an inefficient system with low throughput rates and high drop-out rates can be costly to everyone. The Human Science Research Council (HSRC) in South Africa conducted a study that revealed that on average, 70% of families of the university drop-outs surveyed were categorised in the low economic status civilian category [17]. These students depend on their parents or guardians to pay their fees and/or supplement what they get from government study grants and subsidies such as NSFAS. When these students struggle to complete their studies, it is clear there will be large student debt accumulation for the students or, alternatively, the government incurs costs without a return on investment. In 2005, the National Treasury of South Africa issued a public statement detailing R4.5 billion lost in grants and subsidies without a commensurate return on investment [17].

Evidence that an inefficient higher education system is costly are also explained in another study conducted questioning the drop-out behaviour of learners in universities, where dropping out is shown to have severe consequences for the individuals involved as well as for the society that finances the cost of service delivery [16]. The authors also went on to highlight that having an understanding of the type of students who are more likely to withdraw is important for maximising resource allocation and graduation rates in institutions of higher learning.

With the value and risks associated with going to university outlined, it is clear we need to develop more advanced systems for identifying vulnerable students than what is currently used, as the expansion of enrolments in South African higher education systems has not been accompanied by a proportional

increase in the number of graduates [6]. If we can identify poor performing students early on in the academic year, we can have enough time to remediate their performance to promote success in their undergraduate degree.

This study aims to explore biographical and enrolment observations as tools to predict student performance at a South African higher education institution. The main question we will be asking in the research is, how can we identify students that are at risk of failure in institutions of higher learning at each year of study, based on their biographical and enrolment observations. A synthetic data-set containing enrolment and biographical observations like grade 12 scores, the year started, age, and others, is modelled using various data mining techniques to predict student success in three categories, namely; "First Year Outcome" (FYO), "Second Year Outcome" (SYO), and "Final Outcome" (FO). The data used in this study was from a recent student risk-status study conducted at a South African university [1].

This study argues that there exists characteristics, attributes, and features in a student profile that can accurately predict the student's performance from the first year of registration until qualifying, providing a contribution that suggests a support and supplementary mechanism to the current university Admission Point Score (APS) system, which has evidently been struggling, generating between 25.7% and 32.2% minimum time (3-year) graduates in South Africa for the academic years 2000 to 2017 [10].

Section II will explore the work done by various authors in predicting student performance, exploring the different attributes they found associated with the success of a student. Section III will present the research methodology, outlining the data description and preprocessing, experiments conducted, and models used for prediction. Section IV presents the results and discussion, followed by the concluding section.

II. Related Work

Predicting student performance is the sum of complex and multifaceted factors that cannot easily be represented by student characteristics discovered via student records alone [6]. This chapter, therefore, explores books, journals, and articles concerned with the prediction of student performance using various approaches in order to discover an efficient approach to predicting student performance. Subsection A will analyse and develop categories for the various factors affecting student performance. Subsection B will briefly discuss the methods used by various authors, outlining the factors used and predictive accuracy in each case.

A. Factors affecting student performance

In the South African setting, access to higher education has consistently improved to cover individuals from different backgrounds. Results from a report published by the Department of Higher Education and Training (2020) point out the increasing numbers of first-time students entering university from 98095 students in the year 2000 to over 150 000 students in the academic year 2017 [10]. With this increase comes the idea that the different factors and observations that describe these students must also be increasing, and hence if we aim to accurately partition and describe students into the successful and unsuccessful group, we ought to explore a wide range of observations about each one of them.

1) Socioeconomic observations as determinants of student success: Here we explore variables that are related to an individual or family's measure of social and economic position relative to others. A study conducted on the relationship between socioeconomic status and academic achievement revealed that a correlation exists between the two measures and comments further that the strength of the relationship between the two variables indicates that socioeconomic factors are positively but weakly correlated with academic performance. However, when aggregated groups (grouped data) are the unit of analysis considered, traditional measures of the socioeconomic status usually correlate strongly enough with academic performance to account for some of the variations in a students' performance [31]. Socioeconomic factors that were considered in the research include but are not limited to, family income, education of parents, occupation of the head of the house, and dwelling value.

Other studies that explored socioeconomic factors and their association with a learner's success report financial support and family characteristics as significant factors in explaining drop-out behaviour in higher education [16]. Socioeconomic and other exogenous factors were also found to be significant predictors of student performance in a study of the determinants of student success conducted at a South African University [6]. While exploring socioeconomic factors can certainly improve model accuracy, the drawback of using these measures as a research tool is that they are not straightforward measures of student quality and hence make student performance prediction an even more complex and multifaceted process [6].

2) Psycho-social factors affecting student performance: This subsection explores variables that are related to measures of the combined effects of a students' social factors, thoughts, and behaviour, on academic performance. A study based on psycho-social factors predicting academic performance found that a learners' psycho-social factors such as academic motivation, self-esteem, perceived stress, academic overload, and help-seeking attitude, predict adjustment and academic performance at a historically disadvantaged University in South Africa [27].

A study aimed at predicting first-year college student success made use of six psycho-social factors to construct a model of college success, where success was based on students achieving their academic goals and overall life satisfaction [14]. Hierarchical regression was applied on the six psycho-social factors, namely, academic self-efficacy (describes the student's belief in their capacity to achieve their goals), organisation and attention to study (a measure of the students' time-management behaviour, planning, and scheduling of their college work), stress and time pressure, involvement with college activity, emotional satisfaction with academics, and

Socioeconomic Factors (SEF)	Psycho-Social Factors (PSF)	Pre & Intra-College Scores (PICS)	Individual Attributes (IA)
Family income	Academic self-efficacy	Mathematics	Age at first year
Parents education	Stress and time pressure	English	Work status
Head of house occupation	Class communication	Admission Point Score	Home language
Dwelling value	College activity participation	Accounting	Home province
Dwelling location (rural/urban)	Organization and attention to study	Economic studies	Home country
Financial support	Sense of loneliness	Statistics major	Interest in sports

Table I: This table develops categories for the different features associated with student performance as discussed in Section II. The features are divided into four groups, namely, "Socioeconomic Factors" (SEF), "Psycho-Social Factors" (PSF), "Pre & Intra-College Scores" (PICS), and "Individual Attributes" (IA).

lastly, communication and participation in classes.

The results from the hierarchical regression support those found in other related work [20], as correlation analyses show significant links between student GPA and the six psycho-social factors, with the strongest links involving academic self-efficacy [14]. Academic self-efficacy, student health, students' optimism, and commitment to remain in school are also shown to be strongly related both directly through student performance, and indirectly through expectations and coping perceptions [9].

3) Factors available for this study and other observations: It is clear the students' social surrounding, mindset, and behaviour are significant factors affecting student performance, however, research is subject to multiple constraints such as data availability and participation rates where survey, questionnaires, or other forms of participation is required from students in order to collect the necessary data. A study exploring first-year college student performance reported poor participation in questionnaires, with as little as 23% of the students completing the handed out questionnaires [9]. Other authors report even offering students bonus credits to increase participation and overall data accuracy [14].

More often than not, these and many more limitations of data and data collection methods result in student performance research being conducted majorly using enrolment and other observations from student enrolment records. In this study, biographical and enrolment observations like pre-college scores, age, majors enrolled for, outcomes from all years of study, and a variety of others, are modelled to predict the performance of students at a research-intensive South African university.

Success has been achieved when predicting student performance using similar observations by multiple authors before the current study. This chapter, therefore, continues by introducing a subsection exploring the different methods used by several authors to predict student performance, and the associated predictive accuracy in each case.

B. Methods for predicting student performance

Various authors have already taken on the task of predicting student performance in institutions of higher learning around the world. Many have completed the task with respectable accuracy, where predictive accuracy (P), is measured as:

$$P = \frac{number\ of\ correct\ predictions}{total\ classified\ instances} \tag{1}$$

Table II summarises and compares the accuracies obtained in the various literature reviewed. Leading the table of accuracy with 84.6% predictive accuracy is a study which involved the use of an Artificial Neural Network (ANN) model for predicting the performance of a sophomore student at the Al-Azhar University of Gaza [2]. The study explored Pre & Intra-College Scores (PICS), Socioeconomic Factors (SEF), and Individual Attributes (IA) as determinants of a learner's success. Other researchers also used a similar set of observations to successfully predict student performance. These works include; a study aimed at identifying students at risk of failure conducted at a South African university [1]; and another which took a statistical and data mining approach to the task of student performance prediction [24]. Both studies explored the use of IA, SEF, and PICS, as predictors of various forms of university student performance through the implementation of classifiers like, naïve Bayes, Decision Tree (C4.5), and Sequential Minimal Optimisation (SMO).

Other related work also explored SEF, IA, and PICS, as predictors of student performance but added to these multiple features from the PSF category [20]. These researchers were successful in their task as accuracy's obtained went as high as 84.30% and 80.40% when using naïve Bayes and Neural Network classifiers respectively.

Focusing on the second column of Table II, some important conclusions can be drawn. Firstly, Individual Attributes (IA), appear to be the most prominent of the four categories of predictor features in the prediction of student performance, being utilized in all the literature reviewed. Psycho-Social

Author(s)	Factors considered	Model used	Predictive Accuracy
Abu-Naser *et al.* (2015) [2]	IA, SEF, and PICS	Neural Networks	84.60%
Osmanbegovic and Suljic (2012) [20]	IA, SEF, PICS, and PSF	Naïve Bayes	84.30%
Ajoodha *et al.*(2020) [5]	IA, SEF, and PICS	Random Forest	82.00%
Osmanbegovic and Suljic (2012) [20]	IA, SEF, PICS, and PSF	Neural Networks	80.40%
Osmanbegovic and Suljic (2012) [20]	IA, SEF, PICS, and PSF	Decision Trees (C4.5)	79.60%
Mayilvaganan and Kalpanadevi (2014) [18]	IA and PICS	Decision Trees (C4.5)	74.70%
Ramesh *et al.* (2013) [24]	IA, SEF, and PICS	Multi-layer Perceptron	72.38%
Abed *et al.* (2020) [1]	IA, SEF, and PICS	Naïve Bayes	69.18%
Abed *et al.* (2020) [1]	IA, SEF, and PICS	SMO	68.56%
Ramesh *et al.* (2013) [24]	IA, SEF, and PICS	Decision Trees (C4.5)	64.88%

Table II: A table comparing the different methods used by various authors to predict student performance and the accuracy achieved in each case. The table also provides the different combinations of features used in each case, based on the feature groupings provided in Table 1.

Factors (PSF) as discussed earlier can prove to be accurate determinants of student performance but as seen in Table II, this category is the list used in the reviewed literature due to data restrictions discussed already. Lastly, the table also highlights the importance of having a varied set of predictor variables, as none of the works above explored only one category of features.

Section II has provided sufficient background for us to continue with our task of introducing an approach to the task of student performance prediction, which involves predicting a learners' progress throughout their academic journey, at a research-intensive South African university.

III. Research Methodology

This research proposes an approach to the task of student performance prediction, which involves the prediction of a learner's outcome from the first year of registration until qualifying in a three year degree. This is done by using machine learning predictive models to deduce the outcomes in three different years of study, namely, first, second, and final year. This section introduces the procedure and system of methods applied in this study. Subsection A will give a description of the data and the techniques utilised to make the raw data suitable for machine learning models. Subsection B outlines the features used to predict the three different class variables as well as the methods applied to arrive at the final set of predictor variables. Subsection C provides brief descriptions of the six machine learning classifiers used for prediction.

A. Data Description and Preprocessing

The data-set used for this study is a synthetic data-set generated using a Bayesian Network. The data-set was adopted from a recent study aimed at identifying learners at risk of failure through machine learning procedures, conducted at a South African university [1]. Conditional independence assumptions were used to convey the relationships between enrolment, socioeconomic, and individual attributes such as the year started, age at first year, home country, and a variety of others.

Three target variables are investigated, namely; "First Year Outcome" (FYO), "Second Year Outcome" (SYO), and "Final Outcome" (FO). FYO and SYO contain two similar possible values: proceed, and failed, were to proceed implies the student met the minimum requirements to proceed to the next year of study, while failed implies the student failed to meet the requirements. FO also contains 2 possible values: qualified, and failed, where qualified represents a student who completes the requirements for their chosen degree, while failed in this variable represents a student who failed to obtain their degree.

Data preprocessing is a crucial step in data mining which includes, data preparation, cleaning, normalization, and data reduction tasks such as, feature selection, instance selection, and discretization [11]. The synthetic data-set generated originally contained 50 000 instances. Three random samples (with no replacement of features or bias to uniform class) of 2000 instances were drawn from the data and three phases of experiments were conducted on each sub-sample relative to the target variable we aim to predict in that sub-sample.

The first phase of experiments performed involved the detection of anomalies or outliers. This was done by evaluating classification results from various machine learning models, in an attempt to detect and remove exceptional instances that present significant deviations from the majority patterns. The second phase was aimed at the prevention of over-fitting. This was done by enforcing the same number of training instances in each class through the repeated application of the Synthetic Minority Oversampling Technique (SMOTE). The third phase of preprocessing experiments conducted is feature selection, where 20 features were selected in each sub-sample. Subsection B provides a summary of the factors found for each

sample and a brief discussion of how we arrived at them.

B. Feature selection

#	Feature	1^{st} Year	2^{nd} Year	Final Year
1	English Home Language			
2	Plan Description			
3	Quintile			
4	Home Province			
5	Year Started			
6	Language			
7	Progress Outcome YOS1			
8	Home country			
9	Aggregate YOS2			
10	Rural or Urban			
11	Second Year Outcome			
12	Age at Third Year			
13	Mathematics Literacy			
14	NBTAL			
15	Age at First Year			
16	Computers			
17	NBTQL			
18	Age at Second Year			
19	Life Orientation			
20	NBTMA			
21	Plan Code			
22	English FAL			
23	Additional Mathematics			
24	Mathematics Major			

Table III: A table presenting the various features used for classification. The table sorts the features according to whether they were used for the prediction of the students' 1^{st} Year Outcome, 2^{nd} Year Outcome, or Final Year Outcome.

During the preprocessing phase of the experiments, feature selection was performed on each of the 3 sub-samples drawn from the synthetic data-set. The contribution of a total of 44 features was evaluated using Information Gain (entropy). Using the entropy values alongside multiple experimentations with different subsets, a total of 20 features were selected for training the various machine learning models in each of our 3 cases.

Table III provides a summary of the features used in all the cases considered in this study based on the colour coding scheme developed in Table I. The features are not arranged according to information gain as there are vast differences in the entropy of most features across the three different target variables. The features were chosen to align with our conclusions from the review of previous work, where we concluded the importance of a varied set of predictor variables. The set of predictor variables used in each case has more than two different categories of factors based on the four categories developed in Table I

C. Classification Models

In this research, six off-the-shelf machine learning predictive models are used to predict the target variable at each of the three defined cases. The models used are: Decision tree (C4.5), naïve Bayes, Random Forests, Sequential Minimal Optimization (SMO), Multinomial Logistic Regression, and Logistic Model Trees (LMT). This section gives a brief description of these classification algorithms.

Decision Tree: A decision tree is a decision support system used to learn a classification function that concludes the value of a dependent variable given the values of the independent variables. This classification technique uses tree-like graph decisions and their possible after-effect, including costs of resources, chance results, and utility [19]. There are different algorithms for generating decision tree, J48 (also known as C4.5), Random Forest, and LMT are the chosen models for the purpose of this study.

The C4.5 algorithm utilizes entropy to build a decision tree based on the ID3 algorithm recursively, where features are selected based on information gain. The C4.5 algorithm applied in this study follows the original structure and implementation [23]. LMT uses the combination of a tree structure and logistic regression models to build a single tree. This is done through employing the LogiBoost algorithm for building the regression functions and using the Classification And Regression Tree (CART) algorithm for pruning. The LMT method utilized in this study follows from the original [15].

Random Forests are a combination of decision tree predictors such that each tree in the generated forest depends on the value of a random vector that also governs the growth of each tree. This algorithm is based on growing or generating an ensemble of decision trees from the training data and letting them vote for the most popular class. This multiple decision tree generating technique has several advantages over other classification algorithms including that, the procedure prevents over-fitting through the Law of Large Numbers, it's relatively robust to outliers and noise in data, and the accuracy achieved is as good as with similar machine learning techniques such as Adaboost, while still training faster than bagging or boosting, where we stack classifiers in a similar fashion [7]. The Random Forest implementation used in this paper is based on the original model [7].

Decision trees offer several benefits to data mining which leads to their use in this study. Some of these benefits are: they can handle a variety of input data (nominal, numeric, and textual), they can be implemented from a variety of platforms, and decision tree algorithms can handle missing values in the data-set [19].

Multinomial Logistic Regression: This model is a simple extension of binary logistic regression, allowing for more than two categories of the outcome variable, used to predict categorical placement or the probability of category membership on a dependent variable based on Maximum Likelihood Estimation (MLE) [28]. Although multinomial logistic regression does not assume normality, linearity, or homoscedasticity, the model does require careful examination of outlying cases and sample size [28]. A simple four-category model of this nature, with one independent variable x_i can be represented as: $\log(\frac{\pi_i^{(s)}}{\pi_i^{(0)}}) = \beta_0^{(s)} + \beta_1^{(s)} x_i$, s = 1,2,3,4. Where; $\beta_0^{(s)}$ and $\beta_1^{(s)}$ are intercept and slope parameters, given probability of being in category s can be denoted by $\pi_i^{(s)}$, and chosen reference category $\pi_i^{(0)}$.

These type of models have been implemented by other authors before the current study [13] [30].

Naïve Bayes Classifier: The naïve Bayes model is a simplified example of Bayesian Networks where learning is achieved with ease by assuming that features in the input data-set are all independent given the classifier variable.

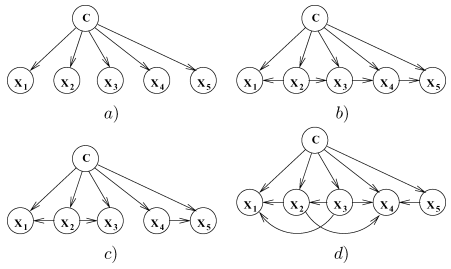

Figure 1: Diagrams (a) to (d) give different examples of Bayesian models representing dependencies among attributes $x_1, x_2, ..., x_5, C$. Where C, is the class variable and the models differ according to the existence of a statistical relationship/dependence between the predictors $(x_1, ..., x_5)$ [11].

Diagram (a) is a depiction of the naïve Bayes model since all the features $(x_1, ..., x_5)$ are conditionally independent given the class. The Naïve Bayes assumption that features are independent given class can be better stated as the distribution: $P(C|X) = \prod_{i=1}^{n} P(x_i|C)$, where $X = (x_1, ..., x_n)$ is a feature vector and C is class. In practice, Naïve Bayes often competes well with more sophisticated classifiers besides its generally poor assumption [25]. The implementation of naïve Bayes in this paper has been implemented before in related studies; [4], [25].

Sequential Minimal Optimization: The SMO is an improved algorithm for training Support Vector Machines which previously required the solution of a large quadratic programming (QP) optimization problem. Traditional training algorithms for SVMs are slow, however, the SMO is much faster as it breaks the large QP problem into a series of the smallest possible QP problems which are solved analytically, avoiding the otherwise numerical QP optimization required. The SMO algorithm implemented in this paper follows the original implementation [21].

D. Prediction and Evaluation

The problem set up in this research is known as a supervised classification problem because the dependent attribute and the values or counting of classes are given. This subsection discusses the various important aspects that are considered and utilized during classification.

1) Evaluation and Validation: To evaluate the effectiveness of each model, a 10-fold cross-validation procedure is applied. This re-sampling technique involves the partitioning of the training data-set, such that a portion of the training data is not seen by the algorithm during training, but is used for model validation. After splitting the data into training and testing set, the training data-set is further split into 10 partitions (folds) where interchangeably 9 folds are used for training and the remaining fold is used for validation until all folds serve as validation fold once.

2) Confusion Matrix: To visualize the performance of the classification algorithms we use a table known as the confusion matrix. The confusion matrix utilized for this study has four important measurement factors as depicted in Table IV.

	Predicted Class +ve	**Predicted Class -ve**
Actual Class +ve	TP	FP
Actual Class -ve	FN	TN

Table IV: A table depicting the structure of the confusion matrix. Negative and positive are depicted by -ve & +ve respectively. Where TP are the true positives, FP the false positives, FN are false negatives, and TN are the true negatives.

3) Accuracy: To determine the performance of the various machine learning models, precision and recall evaluation metrics are obtained from the confusion matrix. Precision (or Confidence in data mining) denotes the proportion of predicted positive cases that are correctly real positive cases. Conversely, Recall denotes the proportion of real positive cases that are correctly predicted positive. Precision and recall are calculated as follows:

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

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

The accuracy follows directly, calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

A similar representation of accuracy has been utilised in various literature reviewed [1], [26]. Other measures of performance evaluated include the Receiver Operating Characteristic (ROC) curve, which plots true positive rate (recall) against the false positive rate (ratio between FP and the total number of negatives). The area under the ROC is of importance as it reflects the probability that prediction is informed versus chance [22]. We want a ROC area above 0.5 as anything below half implies the prediction was guesswork and not informed. Another evaluation measure considered is the F-beta measure (F1 score), which is a measure of a test's accuracy calculated as the weighted harmonic mean of precision and recall, with an optimal value at 1 (meaning perfect precision and recall) and worst value of 0.

IV. Results and Discussion

This section presents the results of the six prediction models discussed in the preceding section. Subsection A gives the accuracy as determined by equation (4), and subsection B presents the confusion matrices obtained, with a discussion of the model accuracy and performance as determined by F-measure and ROC curve.

A. Prediction Outcomes

In this subsection, we present through a table, the predictive accuracy achieved using six different machine learning models to solve our classification problem.

Model used	Predictive Accuracy		
	1st Year Outcome	2nd Year Outcome	Final Year Outcome
Random Forest	94.40%	93.70%	95.45%
LMT	91.90%	91.75%	93.15%
Decision Trees (J48)	87.55%	86.20%	91.45%
Multinomial Logistic	87.80%	86.20%	90.70%
SMO	87.25%	84.45%	89.20%
Naïve Bayes	83.95%	83.40%	84.40%

Table V: Predictive accuracy as calculated by equation (4). After 10-fold cross-validation, all of the models utilized achieved an accuracy above 80%, with Random Forests achieving top accuracy in all three cases considered.

B. Model Performance Evaluation

A study centered around evaluating classification results argued for the use of Recall, Precision, F-Measure, and Receiver Operating Characteristics (ROC) as measures of machine learning model performance [22]. This subsection presents confusion matrices obtained in the three cases of classification, and a brief discussion of performance as observed from F-measure, and ROC.

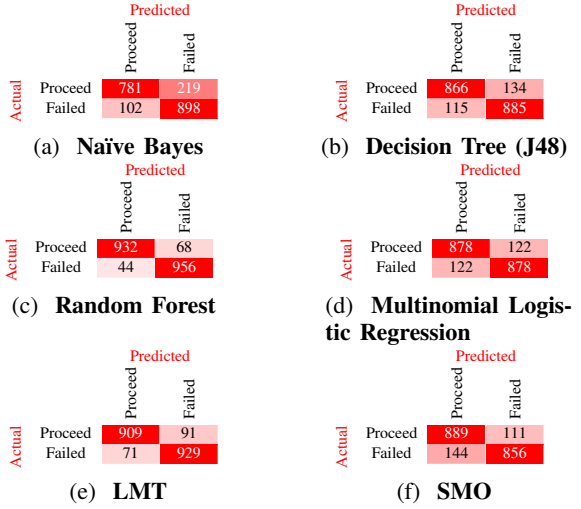

Figure 2: A set of confusion matrices obtained in the prediction of the first-year outcome. Evaluating detailed accuracy by class, we find that, the weighted average of both precision and recall measures is above 0.84 for all six models, furthermore, the F-measure is above 0.83 which is in alignment with our accuracy as depicted in Table V. Area under the ROC curve obtained for all six models also supports the test accuracy as determined by equation (4) and F-measure. The weighted average of the ROC area for each model lies above 0.84 implying our models are making informed predictions and not simply guessing.

We see that the weighted average of both precision and recall measures is above 0.89 for all models except the naïve Bayes, scoring 0.85 and 0.84 respectively. The F-measure also aligns with our high predictive accuracy in table V, as it lies above 0.89 for all cases except naïve Bayes which attains 0.84. The performance of the naïve Bayes model with respect to the other models can be explained by the naive independence assumption it makes. The weighted average of the area under the ROC curve is significant for all six models, as it lies above 0.89 indicating our models were making informed predictions in achieving the high predictive accuracy.

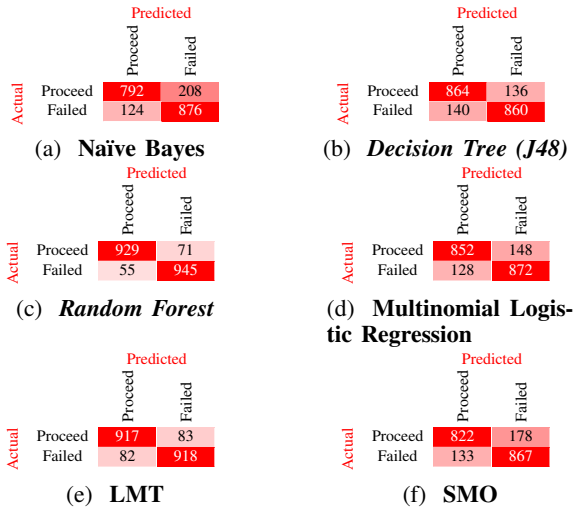

Figure 3: A set of confusion matrices obtained in the prediction of the second-year outcome. Evaluation of detailed accuracy by class reveals that the weighted average of precision and recall measures for all six models lies above 0.83, furthermore, the F-measure of test accuracy is above 0.83 in all cases, which supports our findings in Table V. We also note that the weighted average of the area under the ROC curve for all the six models is in alignment with our accuracy as measured by equation (4) and F-measure. The weighted average of the area under the ROC curve lies above 0.85 for all six models in predicting SYO. This implies our models are not only attaining great predictive accuracy but are making informed predictions to attain this accuracy.

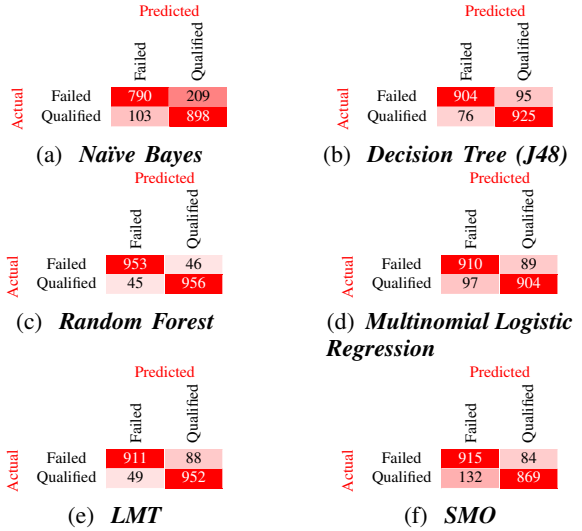

Figure 4: A set of confusion matrices obtained in the prediction of final year outcome.

V. Implications, Conclusion, and Future Work

Enrolment expansion in South African universities has not been accompanied by a proportional increase in the percentage of those who graduate. Various authors have taken on the task to predict student performance as a method to alert students and institutions of the possible trajectory the learners' studies may take.

This paper contributed to the current body of knowledge by introducing an approach that involves the prediction of student performance, at each year of study until qualifying, to provide proactive learner remediation to promote student success. We argue that if we can accurately predict a learner's outcome for the entire academic journey, it can provide an early warning system for those that might struggle and face the various consequences associated with failing or dropping out altogether.

Six machine learning models are utilised to predict first, second, and final year outcomes from a synthetic data-set. After 10-fold cross validation, all six models attained an accuracy above 83% as measured by equation (4) and F-measure of test accuracy. An evaluation of the area under ROC curve also provides constructive feedback as the weighted average of the area under ROC curve for all six models lies above 0.83. The accuracy, F-measure, and ROC curve analysis conducted show that the various classification algorithms can be employed to accurately predict a learners' first, second, and final year outcomes, supporting our initial argument. The significance of this paper is to improve university throughput rates by providing a mechanism to promote student success. To continue with the approach introduced in this study, future work may involve the implementation of the various models on real data, leading to the development of more enhanced and efficient early student performance prediction systems for universities.

VI. Acknowledgements

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

References

[1] T. Abed, R. Ajoodha, and A. Jadhav, "A prediction model to improve student placement at a south african higher education institution," in *2020 International SAUPEC/RobMech/PRASA Conference*. IEEE, 2020, pp. 1–6.

[2] S. S. Abu-Naser, I. S. Zaqout, M. Abu Ghosh, R. R. Atallah, and E. Alajrami, "Predicting student performance using artificial neural network: In the faculty of engineering and information technology," *International journal of hybrid information technology*, 2015.

[3] S. Acid, L. M. de Campos, and J. G. Castellano, "Learning bayesian network classifiers: Searching in a space of partially directed acyclic graphs," *Machine learning*, vol. 59, no. 3, pp. 213–235, 2005.

[4] R. Ajoodha, "Influence modelling and learning between dynamic bayesian networks using score-based structure learning." Wits wirespace, 2018.

[5] R. Ajoodha, A. Jadhav, and S. Dukhan, "Forecasting learner attrition for student success at a south african university," in *Conference of the South African Institute of Computer Scientists and Information Technologists 2020 (SAICSIT '20), September 14-16, 2020, Cape Town, South Africa. ACM, New York, NY, USA, 10 pages.* ACM, 2020.

[6] K. Bokana and D. Tewari, "Determinants of student success at a south african university: An econometric analysis," *The Anthropologist*, vol. 17, no. 1, pp. 259–277, 2014.

[7] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] J. M. Case, D. Marshall, S. McKenna, and D. Mogashana, *Going to university: The influence of higher education on the lives of young South Africans*. African Minds Cape Town, 2018, vol. 3.

[9] M. M. Chemers, L.-t. Hu, and B. F. Garcia, "Academic self-efficacy and first year college student performance and adjustment." *Journal of Educational psychology*, vol. 93, no. 1, p. 55, 2001.

[10] D. DHET Republic of South Africa, "2000 to 2017 fist time entering undergraduate cohort studies for public higher education institutions," ., pp. 16–28, 2020.

[11] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015.

[12] Z. J. Kovacic, "Predicting student success by mining enrolment data," *Research in Higher Education Journal*, vol. 15, p. 1, 2012.

[13] B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 6, pp. 957–968, 2005.

[14] E. J. Krumrei-Mancuso, F. B. Newton, E. Kim, and D. Wilcox, "Psychosocial factors predicting first-year college student success," *Journal of College Student Development*, vol. 54, no. 3, pp. 247–266, 2013.

[15] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine learning*, vol. 59, no. 1-2, pp. 161–205, 2005.

[16] G. Lassibille and L. Gómez, "Why do higher education students drop out? evidence from spain," *Education Economics*, vol. 16, no. 1, pp. 89–105, 2008.

[17] M. Letseka and S. Maile, *High university drop-out rates: A threat to South Africa's future*. Human Sciences Research Council Pretoria, 2008.

[18] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the performance of students academic environment," in *2014 International Conference on Communication and Network Technologies*. IEEE, 2014, pp. 113–118.

[19] B. Neeraj, G. Sharma, R. Bhargava, and M. Mathuria, "Decision tree analysis on j48 algorithm for data mining," *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 6, 2013.

[20] E. Osmanbegovic and M. Suljic, "Data mining approach for predicting student performance," *Economic Review: Journal of Economics and Business*, vol. 10, no. 1, pp. 3–12, 2012.

[21] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998.

[22] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

[23] J. R. Quinlan, "C4. 5: Programs for machine learning," 1993.

[24] V. Ramesh, P.Parkavi, and K.Ramar, "Predicting student performance: A statistical and data mining approach," *INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS*, vol. 63, pp. 975–8887, 02 2013.

[25] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[26] S. Sahu and B. M. Mehtre, "Network intrusion detection system using j48 decision tree," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2015, pp. 2023–2026.

[27] M. Sommer and K. Dumont, "Psycho-social factors predicting academic performance of students at a historically disadvantaged university," *South African Journal of Psychology*, vol. 41, no. 3, pp. 386–395, 2011.

[28] J. Starkweather and A. K. Moske, "Multinomial logistic regression," *Consulted page at September 10th: http://www. unt. edu/rss/class/Jon/Benchmarks/MLR_JDS_Aug2011. pdf*, vol. 29, pp. 2825–2830, 2011.

[29] V. Tinto, "Drop-outs from higher education: A theoretical synthesis of recent research," *Review of Educational Research*, vol. 45, pp. 89–125, 01 1975.

[30] Y. Wang, "A multinomial logistic regression modeling approach for anomaly intrusion detection," *Computers & Security*, vol. 24, no. 8, pp. 662–674, 2005.

[31] K. R. White, "The relation between socioeconomic status and academic achievement." *Psychological bulletin*, vol. 91, no. 3, p. 461, 1982.

[32] J. V. Winters, "Human capital, higher education institutions, and quality of life," *Regional Science and Urban Economics*, vol. 41, no. 5, pp. 446–454, 2011.