Paper 1: Educational Data-mining to Determine Student Success at Higher Education Institutions

Paper 2: A Case Study to Enhance Student Support Initiatives Through Forecasting Student Success in Higher-Education

School of Computer Science & Applied Mathematics
University of the Witwatersrand

Ndou Ndiatenda
828612

Supervised by Dr. Ritesh Ajoodha, and Dr. Ashwini Jadhav

November 30, 2020

A report submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Bachelor of Science with Honours

Abstract

This report presents findings from the research conducted on predicting student performance through biographical and enrollment observations. Chapter 1 comes from a study published with the Institute of Electrical and Electronics Engineers (IEEE). The paper formed part of the IMITEC conference 2020 presenting an educational data-mining approach to the task of student performance prediction, involving the prediction of a learner's end-of-year outcome from the first year of registration until qualifying to graduate in a three year degree. Chapter 2 is an extension of the conference paper presented in the first chapter. The paper presented in the second chapter is stil under review for publication with the Advances in Science, Technology and Engineering Systems Journal (ASTESJ).

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

School of Computer Science and Applied Mathematics

Declaration
2020

I, NDIATENDA NDOU, (Student number: 828612)

am a student registered for Introduction to Research Methods in 2020.

This declaration applies to the Research Report document of Introduction to Research Methods.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand, Johannesburg may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature: _____ Date: 30/11/20

Acknowledgements

My sincere acknowledgements go to my research project supervisors, Dr. Ritesh Ajoodha and Dr. Ashwini Jadhav. This work progressed with ease under your guidance.

Contents

Preface

Abstract i
Acknowledgements iii
Table of Contents iv

1 Educational Data-mining to Determine Student Success at Higher Education Institutions **2**

2 A Case Study to Enhance Student Support Initiatives Through Forecasting Student Success in Higher-Education **11**

*

Chapter 1

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

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 [9]. 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 [33]. 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 [9].

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 [18]. 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 [18].

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 [17]. 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 [7]. 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 [2].

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 [11].

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 [7]. 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 [11]. 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 [32]. 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 [17]. 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 [7]. 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 [7].

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 [28].

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 [15]. 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 [21], as correlation analyses show significant links between student GPA and the six psycho-social factors, with the strongest links involving academic self-efficacy [15]. 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 [10].

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 [10]. Other authors report even offering students bonus credits to increase participation and overall data accuracy [15].

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 [3]. 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 [2]; and another which took a statistical and data mining approach to the task of student performance prediction [25]. 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 [21]. 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) [3]	IA, SEF, and PICS	Neural Networks	84.60%
Osmanbegovic and Suljic (2012) [21]	IA, SEF, PICS, and PSF	Naïve Bayes	84.30%
Ajoodha *et al.*(2020) [6]	IA, SEF, and PICS	Random Forest	82.00%
Osmanbegovic and Suljic (2012) [21]	IA, SEF, PICS, and PSF	Neural Networks	80.40%
Osmanbegovic and Suljic (2012) [21]	IA, SEF, PICS, and PSF	Decision Trees (C4.5)	79.60%
Mayilvaganan and Kalpanadevi (2014) [19]	IA and PICS	Decision Trees (C4.5)	74.70%
Ramesh *et al.* (2013) [25]	IA, SEF, and PICS	Multi-layer Perceptron	72.38%
Abed *et al.* (2020) [2]	IA, SEF, and PICS	Naïve Bayes	69.18%
Abed *et al.* (2020) [2]	IA, SEF, and PICS	SMO	68.56%
Ramesh *et al.* (2013) [25]	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 [2]. 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 [12]. 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 [20]. 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 [24]. 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 [16].

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 [8]. The Random Forest implementation used in this paper is based on the original model [8].

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 [20].

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) [29]. Although multinomial logistic regression does not assume normality, linearity, or homoscedasticity, the model does require careful examination of outlying cases and sample size [29]. 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 [14] [31].

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.

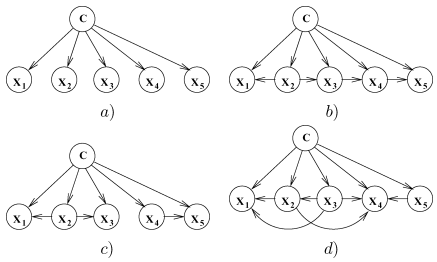


Figure 1: Diagrams (a) to (d) give different examples of Bayesian models representing dependencies among attributes $x_1, x_2, ..., x_5, C$. Where C, is the class variable and the models differ according to the existence of a statistical relationship/dependence between the predictors $(x_1, ..., x_5)$ [12].

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 [26]. The implementation of naïve Bayes in this paper has been implemented before in related studies; [5], [26].

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 [22].

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 [2], [27]. 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 [23]. 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

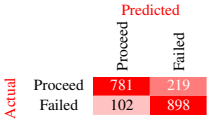
In this subsection, we present through a table, the predictive accuracy achieved using six different machine learning models to solve our classification problem.

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%

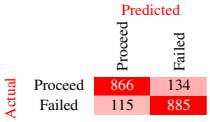
Table V: Predictive accuracy as calculated by equation (4). After 10-fold cross-validation, all of the models utilized achieved an accuracy above 80%, with Random Forests achieving top accuracy in all three cases considered.

B. Model Performance Evaluation

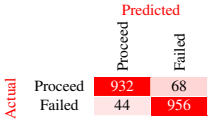
A study centered around evaluating classification results argued for the use of Recall, Precision, F-Measure, and Receiver Operating Characteristics (ROC) as measures of machine learning model performance [23]. This subsection presents confusion matrices obtained in the three cases of classification, and a brief discussion of performance as observed from F-measure, and ROC.



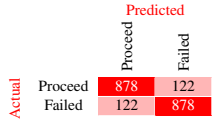
(a) **Naïve Bayes**



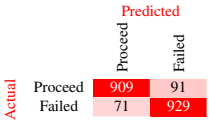
(b) **Decision Tree (J48)**



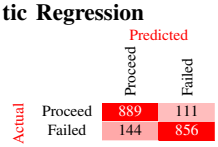
(c) **Random Forest**



(d) **Multinomial Logistic Regression**



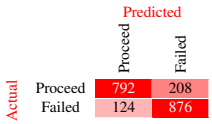
(e) **LMT**



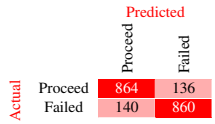
(f) **SMO**

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.

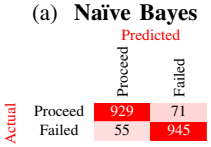
We see that the weighted average of both precision and recall measures is above 0.89 for all models except the naïve Bayes, scoring 0.85 and 0.84 respectively. The F-measure also aligns with our high predictive accuracy in table V, as it lies



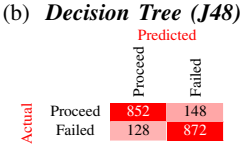
(a) **Naïve Bayes**



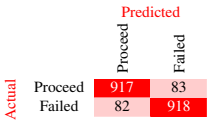
(b) ***Decision Tree (J48)***



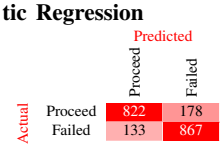
(c) ***Random Forest***



(d) **Multinomial Logistic Regression**

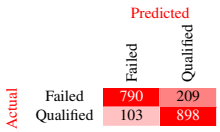


(e) **LMT**

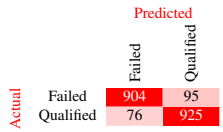


(f) **SMO**

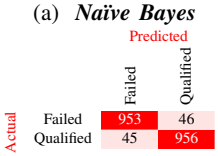
Figure 3: A set of confusion matrices obtained in the prediction of the second-year outcome. Evaluation of detailed accuracy by class reveals that the weighted average of precision and recall measures for all six models lies above 0.83, furthermore, the F-measure of test accuracy is above 0.83 in all 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.



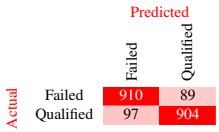
(a) ***Naïve Bayes***



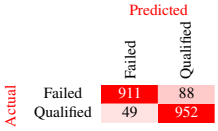
(b) ***Decision Tree (J48)***



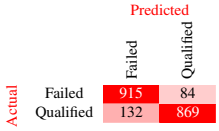
(c) ***Random Forest***



(d) ***Multinomial Logistic Regression***



(e) ***LMT***



(f) ***SMO***

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

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.

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]
[2] 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.
[3] 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.
[4] 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.
[5] R. Ajoodha, "Influence modelling and learning between dynamic bayesian networks using score-based structure learning." Wits wirespace, 2018.
[6] 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.
[7] 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.
[8] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[9] 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.
[10] 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.
[11] D. DHET Republic of South Africa, "2000 to 2017 fist time entering undergraduate cohort studies for public higher education institutions," ., pp. 16–28, 2020.
[12] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015.
[13] Z. J. Kovacic, "Predicting student success by mining enrolment data," *Research in Higher Education Journal*, vol. 15, p. 1, 2012.
[14] 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.
[15] 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.
[16] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine learning*, vol. 59, no. 1-2, pp. 161–205, 2005.
[17] 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.
[18] M. Letseka and S. Maile, *High university drop-out rates: A threat to South Africa's future*. Human Sciences Research Council Pretoria, 2008.
[19] 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.
[20] 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.
[21] 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.
[22] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998.
[23] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.
[24] J. R. Quinlan, "C4. 5: Programs for machine learning," 1993.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] 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.
[30] V. Tinto, "Drop-outs from higher education: A theoretical synthesis of recent research," *Review of Educational Research*, vol. 45, pp. 89–125, 01 1975.
[31] Y. Wang, "A multinomial logistic regression modeling approach for anomaly intrusion detection," *Computers & Security*, vol. 24, no. 8, pp. 662–674, 2005.
[32] K. R. White, "The relation between socioeconomic status and academic achievement." *Psychological bulletin*, vol. 91, no. 3, p. 461, 1982.
[33] 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.

Chapter 2

A Case Study to Enhance Student Support Initiatives Through Forecasting Student Success in Higher-Education

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 5, XX-YY (2020)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

A Case Study to Enhance Student Support Initiatives Through Forecasting Student Success in Higher-Education

Ndiatenda Ndou[*,1], Ritesh Ajoodha[1], Ashwini Jadhav[2]

[1]*School of Computer Science and Applied Mathematics, The university of the Witwatersrand, Johannesburg, 2050, South Africa*

[2]*Faculty of Science, The university of the Witwatersrand, Johannesburg, 2050, South Africa*

ARTICLE INFO

Article history:
Received:
Accepted:
Online:

Keywords:
Higher-Learning
Machine learning
Academic Performance
Student Success
Classification
Random Forests
Student Attrition.

ABSTRACT

Enrolment figures have been expanding in South African institutions of higher-learning, however, the expansion has not been accompanied by a proportional increase in the percentage of enrolled learners completing their degrees. In a recent undergraduate-cohort-studies report, the DHET highlight the low percentage of students completing their degrees in the allotted time, having remained between 25.7% and 32.2% for the academic years 2000 to 2017, that is, every year since 2000, more than 67% of the learners enrolled did not complete their degrees in minimum time. In this paper, we set up two prediction tasks aimed at the early-identification of learners that may need academic assistance in order to complete their studies in the allocated time. In the first task we employed six classification models to deduce a learner's end-of-year outcome from the first year of registration until qualifying in a three-year degree. The classification task was a success, with Random Forests attaining top predictive accuracy at 95.45% classifying the "final outcome" variable. In the second task we attempt to predict the time it is most likely to take a student to complete their degree based on enrolment observations. We complete this task by employing six classifiers again to deduce the distribution over four risk profiles set up to represent the length of time taken to graduate. This phase of the study provided three main contributions to the current body of work: (1) an interactive program that can calculate the posterior probability over a student's risk profile, (2) a comparison of the classifiers accuracy in deducing a learner's risk profile, and (3) a ranking of the employed features according to their contribution in correctly classifying the risk profile variable. Random Forests attained the top accuracy in this phase of experiments as well, with an accuracy of 83%.

1 Introduction

The benefits that can be drawn from institutions of higher-learning extend beyond the degree holder. A study conducted on the relationships between the quality of life, human capital, and universities, revealed that valuable consumption amenities that enhance an areas quality of life are positively correlated with both the local level of human capital (measured by proportion of degree holders in an area) and the number of institutions of higher-learning in a region [2].

The higher-education system is one of great benefit to enrolled-individuals, the economy, and society, however, an inefficient system with high dropout rates and low throughput rates carries harsh costs and consequences for the individual student as well as the society financing the cost of service delivery [3]. The South African Human Science Research Council (HSRC) found that on average, 70% of the university drop-outs they surveyed came from families in the "low economic status civilian" category [4]. Students that belong in this category heavily depend on government study grants and subsidies to supplement the funding they receive from their parents or guardians. It is clear that student debt accumulation, or alternatively, costs to the government without a return on investment will be the outcome when these learners struggle and dropout. This was the case in 2005, where the national treasury reported R4.5 billion lost to student grants and subsidies that resulted in no graduates [4]. There is a student attrition problem in South Africa, as the expansion of enrolments has not come with a significant increase in the percentage of students completing their degrees [5].

Noting the value and possible severe-costs associated with in-

[*]Ndiatenda Ndou, The university of the Witwatersrand, Johannesburg, +27 71 168 8461 & ndiatenda.ndou@students.wits.ac.za

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 1: This table introduces categories for the different features associated with student performance as discussed in Section 2. The features are divided into four groups, namely, "Socioeconomic Factors" (SEF), "Psycho-Social Factors" (PSF), "Pre & Intra-College Scores" (PICS), and "Individual Attributes" (IA). It is also important to take note of the colour coding scheme developed here for later reference to the four categories, [1](sic).

stitutions of higher-learning, we see the clear need to explore the activities influencing student success or failure to solve the problem of student attrition and avoid the severe costs and consequences that it brings. Furthermore, we seek to develop advanced systems for the early-identification of vulnerable learners that may benefit from academic support systems.

In this research, we investigate the influence of biographical and enrolment observations on student success. This research was conducted through two published studies referred to as “phases of the current study” throughout this paper [1, 6]. In the first phase, we employ six machine learning models to predict three target variables that describe a learner’s end-of-year outcome, namely, “first-year outcome”, “second-year outcome”, and “final outcome” [1]. The first phase of this research contributes to the current body of work by showing that various classification models can be used to predict a learner’s end-of-year outcome from the first year of registration until qualifying in a three-year degree. We argue that if we can predict the academic trajectory of a student, early-assistance can be provided to students who may perform poor in the future, remediating their performance and promoting student success.

The second phase of this study involved the prediction of a “risk profile” variable, a variable that categorises the time taken by a student to graduate in a three-year degree by four values, namely: “no risk”, where the student completes the degree in three years; “low risk”, where the student completes the degree in more than three years; “medium risk”, where the student fails/drops-out in less than three year; and “high risk”, where the student takes more than three years to drop out [6]. We used six machine learning algorithms to predict the “risk profile” variable for a student based on biographical and enrolment observations. The second phase of this study contributes to the current body of related literature in three ways: (1) a comparison of six different classifiers in predicting the risk profile of a learner; (2) a ranking of the features employed according to their contribution when deducing the “risk profile”; and (3) an interactive program which uses Random forests classifier to deduce the distribution over a learner’s risk profile. The contribution made by this research implies institutions of higher-learning can use machine learning techniques for the early-identification of learners that may benefit from academic assistance initiatives.

This paper continues with Section 2 which presents the background knowledge around the problem. We then introduce the procedure and system of methods applied in Section 3, followed by the results in Section 4. The work is concluded on Section 5 and we close this study with ideas of future work in Section 6.

2 Related Work

Predicting student performance is a multifaceted task that cannot be easily completed using attributes discovered in student enrolment records alone [5]. We therefore, set-out to explore the various factors influencing student performance, aiming to discover and develop an efficient methodology for solving the problem set up in this research. We begin this chapter with the discussion and grouping of the factors influencing student performance, followed by a brief presentation of the conceptual framework adopted for feature selection in this research, and we close the chapter with a comparison of the various methods of predicting student performance.

2.1 Factors affecting student performance

The South African Department of Higher Education and Training (DHET) released a report with results that portray an increasing number of first-time-entering university students, from 98095 students in the year 2000 to over 150000 students in the 2017 academic year [7]. This significant increase brings the idea that the various attributes that describe a university student must be more varied now than ever before, as more learners from different regions are now enrolling for degrees. To accurately early-identify a member of a given group of learners as successful or unsuccessful in their studies, we must explore a wide range attributes that describe these students, so that our decision is informed.

2.1.1 Socio-Economic Observations Determining Student Success

To determine an individual or their family’s measure of social and economic position in the population, we explore the category of socio-economic attributes as determinants of student success. Studies have shown that a correlation exists between socio-economic status and academic achievement, furthermore, traditional measures of socio-economic status have been revealed to usually correlate strong enough with academic achievement to account for variations

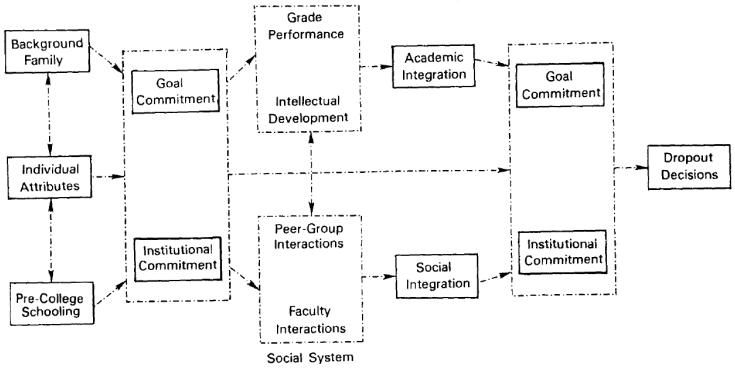


Figure 1: The graphical representation of the framework adopted to determine student success, [6](sic).

in a learner's performance [8]. A study of the determinants of student success also concludes that socio-economic factors contribute significantly when predicting a learner's performance [5]. The socio-economic attributes investigated in the literature we reviewed include; parents' education, dwelling value, family income, and parent/guardian's occupation. Family characteristics and financial support also form part of socio-economic factors revealed to explain student performance, specifically drop-out behaviour in university [3].

2.1.2 *Psycho-Social Factors Affecting Student Performance*

We seek to find more categories of factors that account for variations in a learner's academic performance. We therefore, explore the combined effects of a student's thoughts, social factors, and general behaviour at university on academic performance. Research conducted at a South African university found that to predict a learner's academic performance and adjustment to higher-education, psycho-social factor's such as; help-seeking attitude, workload, perceived stress, and self-esteem could be utilized as determinants [9].

Research on college students' performance utilized six psycho-social factors to predict first-year college student success [10]. The six psycho-social attributes used for prediction were; stress and time pressure levels, communication/participation in class, academic self-efficacy (a learner's belief in their ability to succeed academically), attention to study (measures time-management, planning, and scheduling behaviour), stress levels, and lastly, emotional satisfaction with academics. It was revealed that a strong correlation exists between the six psycho-social factors and a learner's GPA, these findings align with other related work [11]. A separate study also discovered that academic self-efficacy, student's optimism, commitment to schooling, and student health account for some of the variations in a student's performance, expectations and coping perceptions [12].

2.1.3 *Factors Available For This Study*

The review literature centred around factors affecting student performance revealed that a learner's mindset, social surrounding, and behaviour, explain significant variations in the performance of a university learner. In this research, student biographical and enrolment observations such as, majors enrolled for, age, pre-college scores, province and country of origin, are available to build machine learning models for the prediction of academic performance. We continue this chapter by introducing the framework we adopted for the rationale behind predicting academic performance from biographical and enrolment observations.

2.2 *Conceptual Framework*

We use the conceptual framework depicted in Figure 1 as a logical basis to predict the academic performance of a learner from biographical and enrolment observations [13]. The framework develops three categories of attributes contributing to student attrition (drop-out behaviour), namely, background attributes, individual attributes, and pre-college scores. The study reveals that these factors together influence a learner's goal and institutional commitment, which in turn contributes to the drop-out decision of a student via academic and social integration [13]. In this research, we partition background attributes further into socio-economic and psycho-social factors. Table 1 presents the grouping of features discovered to explain variations in academic performance during the review of related literature conducted in this section.

2.3 *Methods For Predicting Student Performance*

Author(s)	Factors considered	Model used	Predictive Accuracy
Abu-Naser *et al.* (2015)[14]	IA, SEF, and PICS	Neural Networks	84.60%
Osmanbegovic and Suljic (2012)[11]	IA, SEF, PICS, and PSF	Naïve Bayes	84.30%
Osmanbegovic and Suljic (2012)[11]	IA, SEF, PICS, and PSF	Neural Networks	80.40%
Osmanbegovic and Suljic (2012)[11]	IA, SEF, PICS, and PSF	Decision Trees (C4.5)	79.60%
Mayilvaganan and Kalpanadevi (2014)[15]	IA and PICS	Decision Trees (C4.5)	74.70%
Ramesh *et al.* (2013)[16]	IA, SEF, and PICS	Multi-layer Perceptron	72.38%
Abed *et al.* (2020)[17]	IA, SEF, and PICS	Naïve Bayes	69.18%
Abed *et al.* (2020)[17]	IA, SEF, and PICS	SMO	68.56%
Ramesh *et al.* (2013)[16]	IA, SEF, and PICS	Decision Trees (C4.5)	64.88%

Table 2: The table compares the accuracy achieved by various authors using different machine-learning models to predict academic performance in each case. The table also illustrates for each case, the combination of features used as predictors, based on the feature categories developed in Table 1, [1] (sic).

Various authors have already accurately predicted student performance utilizing machine learning. Table 2 compares the accuracies obtained in the various literature reviewed in this research. Acquiring the top accuracy position is a research based on the prediction

of the success of a second-year student using an Artificial Neural Network (ANN) model [14]. The study uses individual attributes (IA), pre & intra-college scores, and socio-economic factors as predictors, these inputs are utilized for almost every result presented in the table [11, 17, 16].

3 Research Methodology

This research proposes an approach to the task of university-learner-performance prediction involving the prediction of a learner's end-of-year outcome from the first year they register for a three-year degree until qualifying to graduate. We complete this task by employing six different machine learning algorithms to deduce the outcomes in the first, second, and final year of study in a South African university. This study extends further by attempting to predict a learner's risk-profile based on the time it will take the student to complete a three-year degree.

In this section, we introduce the procedure and system of methods implemented for the purpose of this study. The study incorporates two phases and thus, we begin by giving a description of the two phases, followed by subsection 3.2 giving a description of the data-sets. We present the feature selection technique in subsection 3.3 and follow this by a brief description of the machine learning models we used, closing the section with methods of evaluating and validating our results.

3.1 Phases of the Study

This study was conducted in two phases. The first phase, named, "Preliminary Phase", involved generating preliminary results on a synthetic dataset. In this phase we employed six different machine learning models, namely, Decision tree (C4.5), Logistic Model Trees (LMT), Multinomial Logistic Regression, naïve Bayes, Sequential Minimal Optimization (SMO), and Random forests. The purpose of the first phase was to reveal that machine learning models can be utilized for the early prediction of a learner's end-of-year outcome from the first year of registration until qualifying in a three-year degree, based on biographical and enrolment observations.

The second phase of this study, named, "Post-preliminary Phase", involved the prediction of the distribution over several risk profiles that describe the time it will take for a university student to complete a three-year degree. This phase is performed on a real dataset which the synthetic dataset in the first phase was modelled to resemble. In this phase we employed six different machine learning models, namely, Decision tree (C4.5), Linear Logistic Regression, Support Vector Machines (SVM), naïve Bayes, and Random forests. This phase provides an interactive program which calculates the posterior probability over a learner's "risk profile" as the main contribution o this study, and a ranking of features through Information Gain Ranking (IGR), to determine the features most contributing to student performance.

3.2 Data Description and Pre-Processing

Two sets of data were utilized in this study. We therefore, split the description of the datasets, starting with Subsection 3.2.1 which gives the synthetic dataset description, and Subsection 3.2.2 giving a description of the real dataset.

3.2.1 The Synthetic Dataset Description

The dataset used for the preliminary phase of this study is a synthetic dataset generated using Bayesian Network. The dataset was adopted from a recent prediction modelling study aimed at improving student placement at a South African university [17]. In this dataset, conditional independence assumptions were implemented to portray the relationships that exist between enrolment, socioeconomic, and individual attributes found in student records.

Three target variables are investigated in the preliminary phase, namely, "First Year Outcome" (FYO), "Second Year Outcome" (SYO), and "Final Outcome" (FO). The SYO and FYO variables contain two similar possible values: "proceed", and "failed", were proceed is the outcome for a student who met the requirements to proceed to the next year of study, and failed implies the student failed to meet the minimum requirements to proceed. The FO variable also has two possible values: "qualified", and "failed", were qualified implies the learner met the minimum requirements to graduate in a three-year degree, and failed implies the student failed to meet the requirements to graduate.

Data pre-processing is a crucial step when employing machine learning models. The pre-processing task incorporates, data preparation, data cleaning, data normalization, and data reduction tasks [18]. The synthetic dataset originally contained 50 000 instances. Three random samples (without feature replacement or bias to uniform class) containing 2000 instances were drawn from the raw dataset and several experiments were conducted to make the samples more suitable for our machine learning models. The first set of experiments focused on the detection and removal of outliers or anomalies. This involved the evaluation of classification results from different machine learning classifiers in an attempt to detect and remove instances that display significant deviations from the majority patterns. The second set of experiments conducted aimed to prevent over-fitting. This was done by the implementation of Synthetic Minority Oversampling Technique (SMOTE) which enforced an equal number of training instances for each value in the class variable. The third set of experiments conducted in the preliminary phase is feature selection, where 20 features were selected from each sample based on Information Gain Ranking (IGR) criterion. Subsection 3.3 presents the set of features selected from each sample and a discussion of how we arrived at this set.

3.2.2 The Real Dataset Description

The real data utilized for this study is from a research-intensive university in South Africa. The dataset composes of enrolment and biographical observations of learners enrolled in the faculty of science at the university, from the year 2008 to the year 2018.

The target variable for the post-preliminary phase of this study is "Risk Profile", a nominal variable that tells us how long it will take for an enrolled student to complete a three-year degree at a South African university. The risk-profile variable has four possible values, namely; "no risk", where the student completes their degree in 3 years (the minimum allotted time); "low risk", where

the student completes the three-year degree in more than 3 years; "medium risk", where the student fails to complete the three-year degree before the end of three years (student drops-out in less than 3 years' time); and "high risk", where the student fails to complete the three-year degree after exceeding the 3-years period (student drops-out after exceeding the allotted time).

#	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 3: A table providing the various features selected for input into the employed classifiers. The table groups the features according to whether they were used for the prediction of the learner's 1^{st} year outcome, 2^{nd} year outcome, or final year outcome, [1] (sic).

Since our aim is to perform an accurate classification task on the dataset, pre-processing procedures had to be carried out before training the chosen classifiers. We performed the same experiments as those performed on the synthetic dataset to detect and remove outliers. We then employed a sampling procedure similar to the one we implemented in the preliminary phase of this study. Along with applying SMOTE, a random sample (with no replacement of features or bias to uniform class) of 200 instances was drawn from the original dataset. In Subsection 3.3 we present the features selected and the procedure followed to select them.

3.3 Feature Selection

In this subsection we present the methodology behind the features selected for both phases of the study. Subsection 3.2.1 provides the features we utilized to generate the preliminary results and Subsection 3.2.2 provides the features selected for the post-preliminary phase of the study.

3.3.1 Features for the Preliminary Phase

To select features for the purpose of predicting the three target variables investigated in this phase of the study we utilized Information Gain Ranking (IGR) criterion, which involves deducing the contribution of each feature when classifying an instance as a value of the class variable. Table 3 presents the features selected in all cases considered for the preliminary phase, based on the colour coding scheme developed in Table 1.

Feature selection was performed on each of the 3 samples drawn from the synthetic dataset. We investigated the contribution of 44 features using Information Gain (entropy). Through the entropy values and repeated experimentation with different sets of features, a total of 20 features were selected to predict the target variable in each of the three samples.

The features presented in Table 3 are not arranged according to IGR as there are significant differences in the entropy of most features across predicting the three target variables. The features selected align with our findings from the review of previous work, and more importantly, the conceptual framework we adopt in this research [13]. This is because for each target variable, there were features selected from each of the three investigated categories of features, namely, background (socio-economic) attributes, individual attributes, and pre-college scores.

3.3.2 Features for the second phase

To select features for the post-preliminary phase of this study, we continued utilizing IGR alongside the conceptual framework investigated in the related work section [13]. From the "individual attributes" category, the following features were selected; the National Benchmark Test (NBT) scores for academic literacy (NBTAL), quantitative literacy (NBTQL), and mathematical literacy (NBTMA). These scores give a measure of the individual learner's proficiency and ability to meet the demands of university-level work. To further determine an enrolled learner's professional career aspirations, we also considered the academic plan selected by the learner. The "plan code", "plan description", and "streamline" (Earth Science, Physical Science, Biological Science, or Mathematical Science) variables were selected for this task.

From the "background and family" category, the following features were selected; the home-country and province of the student, whether the school attended by the leaner is in the urban or rural areas, the quintile of the school attended, and the age of the learner at the first-year of registration. These variables combined give us a description of the learner's socio-economic status.

From the pre-college scores category, scores from the following subjects were considered; Mathematics major, Mathematics Literacy, Additional Mathematics, Physical Science, English Home Language, English First Additional Language, Computer studies, and Life Orientation. We note that in this phase of the study, we did not utilize college or university outcomes as inputs, these outcomes were only utilized in the prediction of target variables set up in the preliminary phase as indicated in Table 3. The features selected for the post-preliminary phase of this study are presented ranked according to entropy in Table 6.

3.4 Classification Models

This study composed of two phases, each involving a classification task. In total, we use nine off-the-shelf machine learning predictive models to perform the classification tasks in this research. The models used are: Random forests, naïve Bayes classifier, Decision tree (C4.5), Logistic Model Trees (LMT), Multinomial Logistic Regression, Linear Logistic Regression, Sequential Minimal Optimization (SMO), Support Vector Machines (SVMs), and the K-Star (K*) instance-based classifier. We continue this subsection by giving a brief description of the selected models.

Sequential Minimal Optimization: SMO is an algorithm derived for the improved (speed) training of Support Vector Machines. SVMs previously required that a large quadratic-programming problem be solved in their implementation. Traditional algorithms are slow when training SVMs, however, SMO completes the task much faster by breaking the large quadratic programming problem into a series of smaller problems which are then solved by analytical methods, avoiding the lengthy numerical optimization required. The SMO algorithm we use for the purpose of this study follows the original implementation [19].

Multinomial Logistic Regression: This model derives and extends from the binary logistic regression. It does so by allowing categories of the outcome variable to exceed two. This algorithm utilizes the Maximum Likelihood Estimation (MLE) to predict categorical placement or the probability of category membership on a dependent variable [20]. Multinomial Logistic Regression requires the careful detection and removal of outliers for accurate results, however, the model does not assume or require linearity, homoscedasticity, or normality [20]. An example of a four-category model of this nature, with one independent variable x_i can be given by: $\log(\frac{\pi_i^{(s)}}{\pi_i^{(0)}}) = \eta_0^{(s)} + \eta_1^{(s)} x_i$, s = 1,2,3,4. Where; $\eta_0^{(s)}$ and $\eta_1^{(s)}$ are the slope and intercept respectively, given the probability of category membership in "s" can be denoted by $\pi_i^{(s)}$, and the selected reference category by $\pi_i^{(0)}$. The Multinomial Logistic Regression implementation in this research follows the implementation by other authors before the current study [21, 22].

Naïve Bayes Classifier: The naïve Bayes model is a simplified example of Bayesian Networks. The model achieves learning with ease by assuming that features employed are all independent given the class variable.

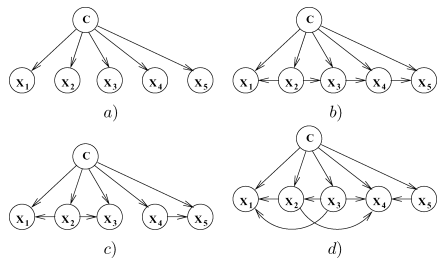


Figure 2: The diagrams (a) to (d) illustrate various examples of a Bayesian network model. The arrows travelling between nodes represent conditional dependencies among the features $x_1, x_2, ..., x_5, C$. Where C, represents the class variable and the models differ based on the existence of a statistical dependence between the predictors $(x_1, ..., x_5)$ [18].

The diagram (a) in Figure 2 is an illustration of the naïve Bayes model. This is because the features, $x_1, ..., x_5$ are conditionally independent given the class variable, C. The naïve Bayes independence assumption can be 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 the class variable. In application, naïve Bayes often performs well when compared to more sophisticated classifiers, although it makes a generally poor assumption [23]. The naïve Bayes model implementation in this research follows other similar implementations in related studies explored; [24, 23].

K* Instance-Based Classifier: The K-Star algorithm classifies instances using training instances that are similar to them alongside a distance function that is based on entropy. The use of an entropy based function provides consistency in the classification of instances in our experiments that may be real-valued or symbolic.The K* implementation utilized for the purpose of this research followed the implementation by [25].

Support Vector Machines: The (SVMs) classification model separates classes of the training data with a hyper-plane. The test instances then get mapped on the same space with their prediction based on the side of the hyper-plane they belong after splitting. This task is performed by incorporating the training dataset into a binary linear classifier that is non-probabilistic. SVMs can be scaled through the one-versus-all partitioning, for various types of classification problems including high-dimensional and nonlinear classification tasks. The SVM model implementation in this research follows the implementation in other related work [26, 27].

Linear Logistic Regression: The Linear Logistic Regression model utilizes additive logistic regression with simple regression functions as base learners of the algorithm [28].The implementation of this model followed in this research follows that of related work conducted in the past [29, 30].

Decision Tree: This model uses a decision support system to build a classification function that predicts the value of a dependent variable given the values of independent variables, through tree-like graph decisions and their possible after-effect, including costs of resources, chance results, and utility [31]. There are different algorithms for generating decision trees; C4.5, Random forest, and LMT are the tree-models selected for the purpose of this research.

The C4.5 algorithm uses information gain to build a decision tree, selecting features based on entropy and utilizing the ID3 algorithm recursively to build the tree. We follow the original structure and implementation of the C4.5 algorithm in this study [32].

LMT builds a single tree from a combination of logistic regression models and a tree structure. This model accomplishes the combination by using the Classification and Regression Tree (CART) algorithm to prune after building the regression functions through the LogiBoost algorithm. The LMT method used in this study follows from the original implementation [29].

Random Forests are a combination of decision tree predictors dependent on the value of a random vector, where the value also governs the growth of each tree in the generated forest. This algorithm involves utilizing the training data to generate an ensemble of decision trees and allowing them to decide on the most-popular class. Implementing this technique has several advantages, including that, the procedure abides by the Law of Large Numbers to prevent over-fitting, it is relatively robust to noise or outliers in

data, and the model achieves accuracy as good as similar techniques such as, Adaboost, Bagging, and Boosting, while still training faster than them [33]. The Random forest model implementation in this research is based on the original model [33].

3.5 Prediction and Evaluation

In this research, the dataset contains the dependent attribute and the values of the classes are known, we therefore have a classification task set up. This subsection provides the measures and techniques utilized to complete this task.

3.5.1 Evaluation and Validation

10-fold cross-validation procedure is applied to evaluate each model employed in this research. In this validation procedure, the training dataset is partitioned such that a portion (testing data) of it is not provided to the algorithm during training, but is used for the validation. The partition remaining for training is further split into 10 partitions (folds). Interchangeably, each of the 10-folds serve for validation while the remaining 9 are used for training until all 10-folds serve as the validation fold once.

3.5.2 Confusion Matrix

We use confusion matrices to illustrate classification outcomes. Table 4 provides an example of the format of a confusion matrix.

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

Table 4: This table depicts the structure of a confusion matrix. We denote "negative" and "positive" by -ve & +ve respectively. Where TP are the true positives (correctly classified positives), FP the false positives, FN are false negatives, and TN are the true negatives (correctly classified negatives).

3.5.3 Accuracy

We extract precision and recall metrics from the confusion matrix in order to measure the accuracy of the employed machine learning models. "Precision" represents the correctly real-positive proportion of predicted positives, while "Recall" represents the correctly predicted proportion of the real positives. We calculate precision and recall as follows:

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

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

The accuracy follows directly, calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

This representation of accuracy has been used in other related work [17], [34]. Other measures of accuracy explored include the Receiver Operating Characteristic (ROC) curve. This curve plots recall (the true positive rate) against the false positive rate (ratio between FP and the total number of negatives), and the area under the ROC reflects the probability that prediction is informed versus chance [35]. We desire the ROC-area to lie above 0.5, anything below 0.5 implies the prediction was guesswork and not informed. Another accuracy measure utilized is the F-beta measure (F1 score), which calculates a test accuracy as the weighted harmonic mean of precision and recall. The optimal value for the F-measure is 1 (indicating perfect precision and recall), and the worst value is 0.

4 Results and Discussion

This study was conducted in two phases. The first phase involved generating preliminary results on a synthetic data-set, while in the second phase, a similar set of experiments are performed on a real data-set, leading to conclusions and implications about the performance of the trained machine learning models in classifying the problem at hand, furthermore, the results drawn from the second phase provide a ranking of the employed features according to entropy, together with an interactive program which calculates the posterior probability over the students' risk profile so that support initiatives and programs can be focused on them.

4.1 Preliminary Results

This Subsection presents the results of six of the nine prediction models discussed in Section 3, namely; Decision tree (C4.5), Logistic Model Trees (LMT), naïve Bayes Classifier, Sequential Minimal Optimization (SMO), Multinomial Logistic Regression, and Random Forests. We present first the prediction outcomes through a table comparing predictive accuracy of the models as determined by Equation 3. This will be followed by an evaluation of the model's performance through F-measure and Receiver Operating Characteristic (ROC) curve.

4.1.1 Prediction Outcomes

Six different machine learning models were utilized to solve our classification problem. The predictive accuracy achieved by each model is recorded and presented in the Table 5.

	Predictive Accuracy		
Model used	1^{st} Year Outcome	2^{nd} 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 5: Predictive accuracy as calculated by Equation 3. 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, [1](sic).

4.1.2 *Model Performance Evaluation*

A study based on evaluating classification results argued for the use of Precision, Recall, F-Measure, and the ROC curve, as accurate measures of a machine learning model performance [35]. We utilize these measures to evaluate results presented by Figure 3, 4, and 5.

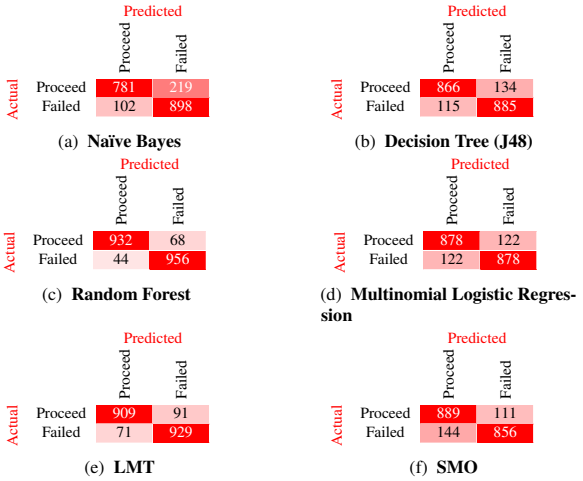


Figure 3: The set of confusion matrices resulting from the prediction of first-year outcome. Evaluation of the accuracy by class reveals that the weighted average of both precision and recall lies above 0.84 for all six models trained in our study. Further observations reveal that the f-measure of accuracy is more than 0.83 for all models, this value aligns with our accuracy as determined by Equation 3 in the Table 5. The test accuracy obtained in the table is further supported by the ROC curve obtained for all six models, as the weighted average of the ROC area for each model is more than 0.84 implying the models trained were making informed decisions in classifying the problem and not simply guessing.

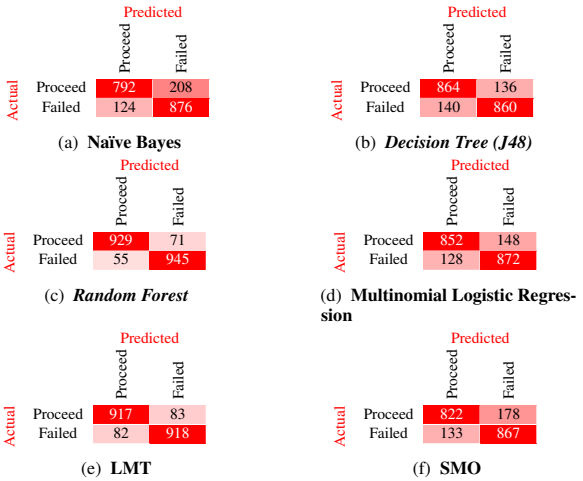


Figure 4: The confusion matrices obtained when classifying the second-year outcome variable. When we evaluate the detailed accuracy by class for each model, we find that the weighted average of both precision and recall measures is above 0.83, furthermore, the f-measure of test accuracy aligns with our findings in Table 5 with a value of more than 0.83 for all six models considered. The ROC curve obtained for all six models also aligns with our accuracy as determined by f-measure of test accuracy and Equation 3, as the weighted average of the area under the ROC curve lies above 0.85 implying the models trained are not attaining the great predictive accuracy through guess-work, but they are making informed predictions.

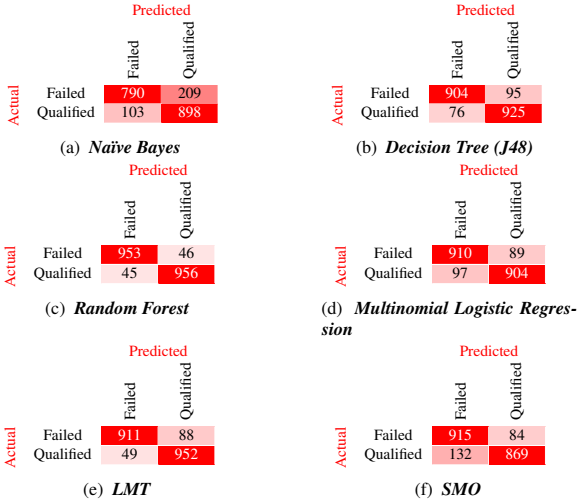


Figure 5: Confusion matrices obtained when classifying the third-year outcome variable with the six trained machine learning models. Evaluating the detailed accuracy by class results, we find that the weighted average of both precision and recall measures is more than 0.89 for all models except the naïve Bayes classifier which scored 0.85 for precision, and 0.84 for recall. The underperformance of the naïve Bayes model with respect to other trained classifiers can be explained by the naïve feature-independence assumption it makes. Results also show that the weighted average of the area under the ROC curve for each model is in alignment with our test accuracy measures, as it lies above 0.89, implying each model is making informed predications and not simply guessing the outcome.

4.2 *Main Results (Post-preliminary)*

The preliminary phase of this study revealed that we can utilize machine-learning models to accurately predict a learner's outcome from the first year of registration until qualifying in a three-year degree. The preliminary results confirmed our initial hypothesis but furthermore, revealed the kind of model we should utilize with such a wide variety of models available for use, but few fitted to the problem set up in this study. Evaluating the preliminary results, we note that Random Forests achieved top accuracy and performance as measured by all our model performance and accuracy evaluators, across all three test cases.

In this subsection, we present the results obtained when utilizing machine learning models to classify a learner into the four risk profiles ("No Risk", "Low Risk", "Medium Risk", and "High Risk") defined in Section 3, as the preliminary phase has confirmed this task can be completed. The classification problem set up in this phase is parallel to that in the prelim-phase in several ways, as the preliminary phase utilizes a synthetic data-set modelled to resemble the relationships that exist within the student enrolment data utilized for the second phase.

Subsection 4.2.1 presents the selection and ranking of features utilized, we follow this by a presentation of the classification outcomes and close the section with the presentation of an interactive program that can be utilized to calculate the posterior probability over a student's risk profile.

4.2.1 *Selection and Ranking of features*

We selected 20 features to predict the class variable. The features were selected using Information Gain Ranking (IGR) to deduce

the contribution of each feature in classifying the instances. The feature selection findings are illustrated through Table 6 with three columns below. The first column determines the rank of the features among the input set, the second column gives the amount of entropy attained by the feature, and the third column has the name of the feature associated with the ranking and entropy.

Rank	Entropy	Feature Name
1	1.21960228	PlanCode
2	1.15086266	PlanDescription
3	0.59886383	Streamline
4	0.29582771	Year Started
5	0.20836689	AgeatFirstYear
6	0.18695721	SchoolQuintile
7	0.14234042	MathematicsMatricMajor
8	0.12166049	Homeprovince
9	0.06417526	isRuralorUrban
10	0.0568866	LifeOrientation
11	0.04978826	PhysicsChem
12	0.02780914	EnglishFirstLang
13	0.01253064	Homecountry
14	0.00550434	AdditionalMathematics
15	0.00000902	MathematicsMatricLit
16	< 0.00001	NBTAL
17	< 0.00001	NBTMA
18	< 0.00001	NBTQL
19	< 0.00001	ComputerStudies
20	< 0.00001	EnglishFirstAdditional

Table 6: A ranking through information gain (entropy) of the set of features selected to predict the student risk profile, ranked from the most contributing feature to the least contributing feature. The top seven features that are highlighted indicate an entropy greater than 0.1, [6] (sic).

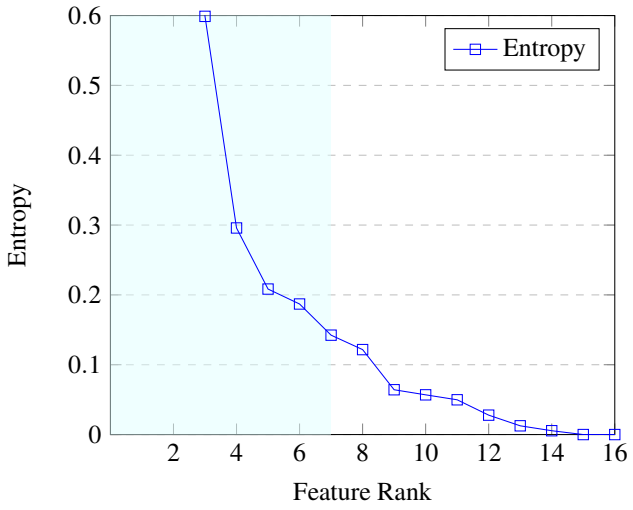


Figure 6: A graphical illustration of how the information gain (entropy) level varies across the chosen feature input set. The x-axis indicates the feature rank, and the y-axis indicates the information gain from utilizing the corresponding feature, [6] (sic).

The use of IGR also has implications on the contribution made by a feature relative to others within the chosen input set of features. We investigate the behavior of entropy as you move between subsequent features and present the findings through a graphical illustration in Figure 6, showing the monotonically decreasing behavior of the entropy function plotted versus rank. We see that the loss in entropy between each subsequent point (feature rank) is logarithmically decreasing, furthermore, we highlight the same top seven features from the IGR Table 6.

The highlighted features on Figure 6 illustrates their relative importance in forecasting the success of a learner. We note that the set of factors most contributing to a learner's success includes, "the plan-code and plan description" (combined, these two variables give a precise description of what the student is studying), "streamline" (mathematical, life, or physical science), "the year started", "school quantile", "age at first-year", and "the student's matric mathematics score".

4.2.2 *Classification Outcomes*

This section presents the results obtained from the classification algorithms trained to predict the class variable (risk profile). Six of the nine classification procedures discussed in Section 3 were employed for the post-preliminary phase of the study: Decision trees (C4.5), naïve Bayes Classifier, Linear Logistic Regression model, Support Vector Machines (SVMs), K*, and Random Forests.

Figure 7 (a) – (f) illustrates the results of each trained classifier after 10-fold cross-validation. Evaluating the performance of each model relative to other models employed, we note that the Random forests classifier attains the highest accuracy (83%) of all the models trained for the post-preliminary phase. This result aligns with our findings from the preliminary phase of the study as Random forests attained the highest accuracy among the selected models for both phases of the study.

We note further that the SVM classification model was revealed to be the least suited for the problem set up in this study. At 52%, SVMs attained the lowest predictive accuracy in this study, furthermore, SVMs took the longest time to train. When discussing training and testing times, it is also important to note that the K-star model took the least time to implement in this study.

Overall, the classification task was a success, with five of the six models employed attaining a predictive accuracy above 75%. Noting that Random forests was the most accurate in predicting the class variable for both phases of this study, this section continues by providing a web application utilizing the Random forests classifier to predict the risk-profile of a learner based on enrolment and academic factors. Severity of misclassified instances was also evaluated and taken into account to determine Random forests as the most suited model for the task, for example, the 27% of "No Risk" instances incorrectly classified by SVM as "Medium Risk" is far more severe and misleading than the misclassification of 5% "No Risk" instances as "Medium Risk" by Random forests classifier.

4.2.3 *Main Contribution of The Study*

In this subsection we provide an interactive program which can calculate the posterior probability over a student's risk profile utilizing the Random forests model employed in Section 4.2.2. This automated system makes predictions about a student's risk of fail-

Actual Risk \ Predicted Risk	No	Low	Med	High
No	171	20	9	0
Low	26	143	25	6
Med	17	22	145	16
High	2	7	16	175

(a) A confusion matrix resulting from the prediction of a student's "Risk Profile" utilizing the **C4.5** classification model. The **C4.5** algorithm achieved **79**% accuracy, furthermore, 634 instances were correctly classified and 166 instances were incorrectly classified by the model.

Actual Risk \ Predicted Risk	No	Low	Med	High
No	159	24	11	6
Low	25	145	14	16
Med	14	21	124	41
High	6	7	12	175

(b) A confusion matrix detailing the performance of the **K*** classifier when predicting a student's "Risk Profile" (class variable). The lazy K-Star algorithm achieved **75**% accuracy, furthermore, 603 instances were correctly classified and 197 instances were incorrectly classified by the model.

Actual Risk \ Predicted Risk	No	Low	Med	High
No	174	16	10	0
Low	31	142	22	5
Med	8	17	160	15
High	6	8	17	169

(c) A confusion matrix detailing the performance of the **naïve Bayes** classifier when predicting a student's "Risk Profile" (class variable). After 10-fold cross-validation the naïve Bayes classifier achieved **80**% accuracy, furthermore, 645 instances were correctly classified and 155 instances were incorrectly classified by the model.

Actual Risk \ Predicted Risk	No	Low	Med	High
No	108	23	55	14
Low	38	80	65	17
Med	29	34	125	12
High	37	26	34	103

(d) A confusion matrix resulting from the prediction of a student's "Risk Profile" utilizing the **SVM** classification algorithm. The SVM algorithm achieved **52**% accuracy, furthermore, 416 instances were correctly classified and 384 instances were incorrectly classified by the model.

Actual Risk \ Predicted Risk	No	Low	Med	High
No	174	17	9	0
Low	23	150	21	6
Med	6	15	161	18
High	5	4	14	177

(e) A confusion matrix resulting from the prediction of a student's "Risk Profile" utilizing **Random Forests** classifier. After 10-fold cross-validation the random forests model achieved **83**% accuracy, furthermore, 662 instances were correctly classified and 138 instances were incorrectly classified by the model.

Actual Risk \ Predicted Risk	No	Low	Med	High
No	168	21	11	0
Low	37	138	19	6
Med	6	22	149	23
High	4	4	18	174

(f) A confusion Matrix detailing the performance of the **linear logistic regression** classifier when applied to a dataset of biographical and enrolment observations. The linear logistic regression model achieved **78**% accuracy, furthermore, 629 instances were correctly classified and 171 instances were incorrectly classified by the model.

Figure 7: A set of confusion matrices obtained when classifying the "Risk Profile" variable. We provide the accuracy of each classification model as determined by Equation (*****) along with a count of the correctly and incorrectly classified instances by each model, [6] (sic).

ure based on the conceptual framework developed in a "drop-out from higher education" study [13]. The conceptual framework connects dropout decision to categories of input features, namely, background (family) attributes, individual attributes, and pre-college scores. This framework is better depicted in the Figure 1 provided under the related work section.

The web application depicted in Figure 8 provides a practical tool which university student support programs can utilize for early detection of learners in need of academic support. We argue that the early detection and assistance of students at risk of failure is likely to lead to improved academic performance and eventually higher pass-rates, which translate to increased throughput rates.

The example depicted in Figure 8 illustrates the calculation of the risk profile of a learner based on biographical and enrolment observation. The program predicts the posterior distribution over the four "Risk Profiles", namely, "No Risk", "Medium Risk", "High Risk", and "Low Risk" using the Random forests classifier. The learner in the example is from an urban quantile 3 school in Kwazulu-Natal South Africa, furthermore, individual attributes are provided as follows; scores of 48%, 50%, and 55% for the quantitative, academic, and mathematical literacy National-Benchmark-Tests (NBT) respectively. The learner also completed pre-college courses with scores of; 50% for both core Mathematics and Life Orientation, and 60% for English Home Language. The output presented at the bottom of Figure 8 illustrates that hypothetically the student is 10% likely to complete their degree in 3 years (No Risk), 50% likely to complete their degree in greater than 3 years (Low Risk), 35% likely to drop-out before the end of 3 years (Medium Risk), and 5% likely to drop-out in greater than 3 year (High Risk). With the output obtained, students support-program-coordinators can then decide what assistance will prove most beneficial to the student in the example as the learner posses a high chance of struggling to complete their degree in the allocated time (50% Low Risk).

5 Implications and Conclusions

The expansion of enrolments in South African universities has not been accompanied by a proportional increase in the percentage of learners graduating. In this study, we took on the task of early prediction of a learner's academic trajectory, aiming at identifying those who may struggle in universities, so that proactive learner

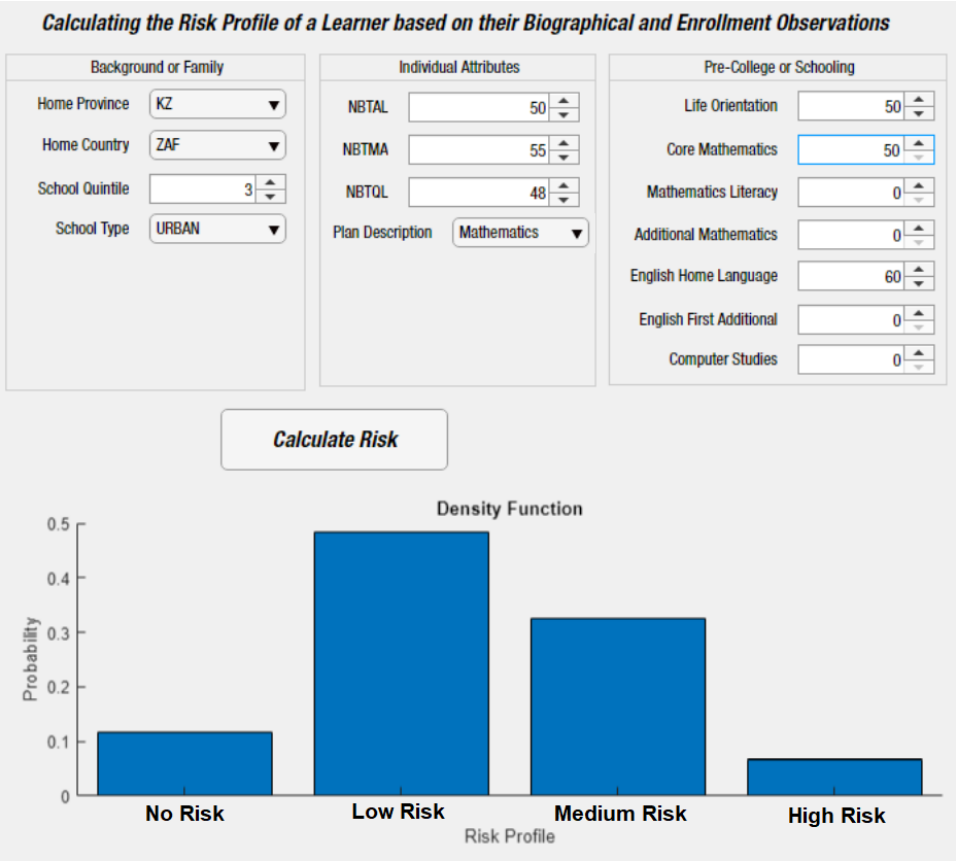


Figure 8: The graphical user interface for the at-risk program [6](sic).

remediation to promote success may be provided to them.

This paper contributes to the current body of knowledge firstly by introducing an approach involving the prediction of a learner's outcome from the first year of registration until qualifying in a three-year degree. We argue for the early prediction of a learner's entire academic trajectory with the aim to detect those who are likely to benefit from student academic support initiatives. We trained six machine learning models to predict first, second, and final year outcomes from a synthetic data-set. After 10-fold cross-validation this task was completed with great success as all six models attained an accuracy above 83%. Furthermore, an evaluation of the F-measure of accuracy and ROC-curve reveal that these models are making informed-accurate decisions and not simply guessing, therefore, leading to our second contribution involving a real data-set from a research-intensive university in South Africa.

The second contribution of this study is a ranking (through entropy) of features according to their contribution in correctly predicting a learner's "risk profile" (class variable). The ranking of features according to entropy reveals which features are stronger determinants of student success relative to others employed and, in this study, we highlight the seven top-ranked features, namely, "plan code", "plan description", "year started", "age at first year", "streamline", "school quantile", and "Matric Mathematics major".

The third and main contribution made by this study is an interactive program which can predict the distribution over a learner's risk profile utilizing biographical and enrolment observations. The interactive program proposed in this paper can be utilized for early identification of university learners who are most likely to benefit from student support initiatives aimed at improving academic performance. The implication of this study is that university learners can be assisted early in their academic journey increasing their chances of success. Furthermore, the early detection and assistance of learners in need of academic support will result in an improved and enriched learning experience beyond what the student would have experienced if support initiatives were implemented after failure has been detected.

6 Future work

To continue with the work done in this paper, future work may involve: (a) incorporating into our models features from categories not considered such as the "psycho-social attributes category", (b) exploring what courses offered in university possess high failure rates and how good the set of features we employed predict success in these courses, or (c) approaching the problem from a different perspective by attempting to predict which courses are students likely to struggle completing so that support initiatives may be focused on the specific courses.

Conflict of Interest The authors declare no conflict of interest.

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

References

[1] Author, "The publisher has been removed for the peer review process," in The publisher has been removed for the peer review process, IEEE, 2020.

[2] J. V. Winters, "Human capital, higher education institutions, and quality of life," Regional Science and Urban Economics, **41**(5), 446–454, 2011.

[3] G. Lassibille, L. Gómez, "Why do higher education students drop out? Evidence from Spain," Education Economics, **16**(1), 89–105, 2008.

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

[5] K. Bokana, D. Tewari, "Determinants of student success at a South African university: An econometric analysis," The Anthropologist, **17**(1), 259–277, 2014.

[6] Author, "The publisher has been removed for the peer review process," in The publisher has been removed for the peer review process, ACM, 2020, doi:https://doi.org/10.1145/3410886.3410973.

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

[8] K. R. White, "The relation between socioeconomic status and academic achievement." Psychological bulletin, **91**(3), 461, 1982.

[9] M. Sommer, K. Dumont, "Psycho-social factors predicting academic performance of students at a historically disadvantaged university," South African Journal of Psychology, **41**(3), 386–395, 2011.

[10] E. J. Krumrei-Mancuso, F. B. Newton, E. Kim, D. Wilcox, "Psychosocial factors predicting first-year college student success," Journal of College Student Development, **54**(3), 247–266, 2013.

[11] E. Osmanbegovic, M. Suljic, "Data mining approach for predicting student performance," Economic Review: Journal of Economics and Business, **10**(1), 3–12, 2012.

[12] M. M. Chemers, L.-t. Hu, B. F. Garcia, "Academic self-efficacy and first year college student performance and adjustment." Journal of Educational psychology, **93**(1), 55, 2001.

[13] V. Tinto, "Drop-Outs From Higher Education: A Theoretical Synthesis of Recent Research," Review of Educational Research, **45**, 89–125, 1975, doi: 10.2307/1170024.

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

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

[16] V. Ramesh, P.Parkavi, K.Ramar, "Predicting Student Performance: A Statistical and Data Mining Approach," INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS, **63**, 975–8887, 2013.

[17] Author, "The publisher has been removed for the peer review process," in The publisher has been removed for the peer review process, 1–6, IEEE, 2020.

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

[19] J. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," Advances in Kernel Methods-Support Vector Learning, **208**, 1998.

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

[21] B. Krishnapuram, L. Carin, M. A. Figueiredo, A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," IEEE transactions on pattern analysis and machine intelligence, **27**(6), 957–968, 2005.

[22] Y. Wang, "A multinomial logistic regression modeling approach for anomaly intrusion detection," Computers & Security, **24**(8), 662–674, 2005.

[23] I. Rish, et al., "An empirical study of the naive Bayes classifier," in IJCAI 2001 workshop on empirical methods in artificial intelligence, volume 3, 41–46, 2001.

[24] R. Ajoodha, Influence Modelling and Learning Between Dynamic Bayesian Networks Using Score-based Structure Learning, University of the Witwatersrand, Faculty of Science, School of Computer Science and Applied Mathematics, 2018.

[25] J. G. Cleary, L. E. Trigg, "K*: An Instance-based Learner Using an Entropic Distance Measure," in 12th International Conference on Machine Learning, 108–114, 1995.

[26] Y. EL-Manzalawy, "WLSVM," 2005, you don't need to include the WLSVM package in the CLASSPATH.

[27] C.-C. Chang, C.-J. Lin, "LIBSVM - A Library for Support Vector Machines," 2001, the Weka classifier works with version 2.82 of LIBSVM.

[28] J. Friedman, T. Hastie, R. Tibshirani, "Additive Logistic Regression: A Statistical View of Boosting," The Annals of Statistics, **28**, 337–407, 2000, doi: 10.1214/aos/1016218223.

[29] N. Landwehr, M. Hall, E. Frank, "Logistic Model Trees," Machine Learning, **59**, 161–205, 2005, doi:10.1007/s10994-005-0466-3.

[30] M. Sumner, E. Frank, M. Hall, "Speeding up Logistic Model Tree Induction," in 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, 675–683, Springer, 2005.

[31] B. Neeraj, G. Sharma, R. Bhargava, M. Mathuria, "Decision tree analysis on j48 algorithm for data mining," Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, **3**(6), 2013.

[32] J. Quinlan, C4.5: Programs for Machine Learning, Ebrary online, Elsevier Science, 2014.

[33] L. Breiman, "Random forests," Machine learning, **45**(1), 5–32, 2001.

[34] S. Sahu, B. M. Mehtre, "Network intrusion detection system using J48 Decision Tree," in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2023–2026, IEEE, 2015.

[35] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," Mach. Learn. Technol., **2**, 2008.