Evaluation of Student Skill-Sets as predictors of Success in Degree Streamlines at the Faculty of Science.

Luyanda Makhoba (834867)

Supervisor:
Dr. Ritesh Ajoodha

University of the Witwatersrand
School of Computer Science and Applied Mathematics

Abstract

This paper presents an approach of using a skill-set based model representation of student academic ability in order to predict their suitability for a science degree streamline (i.e *Biological, Earth, Mathematical and Physical Science*). This approach offers an alternative to conventional university entrance criteria, establishing a means to better advice prospect university students on which degree streamline they are best suited for based on their displayed abilities. This will assist in the measures taken to reduce student attrition at universities. Student skill-sets are composed of pre-college marks (high school leaving results and National Benchmark Test scores) and home language. The composition of skill-sets is based on key student abilities associated with student university performance, these are characterised as follows: *Mathematical Ability, Computer Proficiency, Academic Literacy and Language Communication.*

Through this research we investigate the following :

I. The ranking of key skills that influence the success of students per streamline

II. The ability to determine probability of attrition for each new student by streamline

III. Evaluate and compare best performing predictive models

Findings of the report outline the extent at which skill-sets can be used to predict the students suitability for degree streamline better than the university Admission Point Score method. The best performing prediction model proved to be the K-Star classifier which gave an accuracy of 80.4 %. It is also worth noting that the proposed approach encompasses more logical improvements then the currently existing structure of acceptance at South African universities. This research, conducted at a South African research intensive institute, shows that skill-sets can be used as an alternative to current university entrance requirements as they provide a more holistic view of the students ability.

In this paper We argue that skill-sets, as presented in this study, are capable of predicting student success and attrition better than the current university entrance requirements criteria that uses APS as a means of student acceptance. With the advent of machine learning, this study demonstrates the capabilities of data mining in education as viable solutions to the student attrition problem with skill-sets giving a more holistic representation of the students capabilities making for better classification predictions.

Key Words: *Data mining, Student attrition, classification models, skill-set*

Nomenclature

Abbreviation/Jargon	Explanation
APS	Admission Point Score
AUC	Area Under Curve. An evaluation metrics
Attrition	Student failure or dropout opposed to graduation
College	University, Tertiary Education Institute
False Positive/Negative	Miss-classified class values
IGR	Information Gain Ranking
NBT	National Benchmark Test
ROC Curve	Receiver Operating Characteristics.Illustrates accuracy of model performance
Skill-set	Representation of learner ability
Student	Learner, Any individual enrolled, or to be enrolled in University
SVM	Support Vector Machine. Classification model
True Positive/Negative	Correctly classified values

Table 1: Description of abbreviations and jargon used in this research paper.

Contents

1 Introduction

The current entrance requirements and acceptance criteria used by South African university institutes employ a technique that does not evaluate student suitability for a degree program substantially. This is evident when investigating the attrition rate of students at tertiary institutes. With the use of admission point score (APS) as a means of student acceptance within degree streamlines, it has been found that only roughly 30% of first-time students are able to graduate and obtain their degrees after a five-year period. [1, 2].

In order to aid the student admission process and offer specific intervention support for incoming students, a practical model that better represents the capabilities of students is needed. The APS system works by assigning scores to seven selectively chosen subjects taken by a student, these scores are calculated using on a discretized approach where a subject mark is given a point score if it is within a certain range (taken as intervals of tens). This discretization sees the loss of information and causes student ability misrepresentation. For example a student with a mark of 69 would be characterised the same as a 60 despite being closer to a 70. As an alternative to the flawed conventional APS entrance criteria, we propose a more bespoke model to represent student ability. One that characterises a learners ability as skill-sets in-line with practical university requirements.

The basis of the model approach used in this research was inspired by the student attrition model proposed by Tinto [3]. The model depicts the relationship between college student attrition and other factors of influence such as family background, pre-college schooling and the student's individual attributes. A students grade performance represents their intellectual development which happens in parallel with other social developments based their social integration. After careful consideration, we made adjustments to the original model to form the conceptual framework that this research paper encompasses. The conceptual framework of this study is illustrated in figure 1, it depicts the key notion of how skill-sets of a student are formed and how they essentially determine a student's university success or attrition.

Based on this underlying scheme, we define the composition of each of the skill-sets, each of them incorporates pre-college marks. With this study we investigate the feasibility of using a model composed of a students skill set, and consequently use this to predict their success in a particular Science degree streamline. The four Science streamlines are partitioned as such: *Biological Sciences, Earth Sciences,* *Mathematical Science and Physical Sciences*. The skill sets are composed of high school results, National Benchmark Tests(NBT) and biographical data. The results of this proposed research provide insight into which core skills can be used as predictive identifiers of success in the respective science degrees. In this research the skill-sets used are namely *mathematical ability, Academic Literacy, Communication skills and Computer proficiency*. Figure 4 outlines the composition of the learner skill-sets.

The purpose of this research is to be able to assist students in making decisions on which degree they are most likely to be successful in based on their skill-set characteristics. The insight from this research can also be used by university administration staff to identify at risk students and more specifically identify the additional assistance the student will require based on the skill set characteristic that falls short. Similarly, this approach can be used by bursary and scholarship sponsors who can engage with the student by offering them further support in the either of the skill set domains where the student is low scoring in their respective skill sets. Different forms of support and Interventions can be made that can improve the prospects of students [1].

A range of machine classification models were chosen and trained on the data, namely Decision Trees, Decision Tree Naive Bayes, Multilayer Perceptron, K-star and SVM. We then identified the best performing algorithm, we found this to be the K-Star model with a predictive accuracy of 80.4%. This aspect forms the basis for the first contribution of this research which is the identification of a means to predict student attrition. Other contributions of this research include the rankings of the usefulness of each of the skill-sets when predicting student attrition. We found that learner writing ability significantly has the highest information gain, followed by Mathematical ability. Language and Computer Proficiency were the least ranked so much so that they could be interpreted as interchangeable. The last major contribution of this research is the inception of an application that can be used by students in finding out their probability of success under the different Science streamlines based on their attributes that form the skill-sets representing their displayed abilities.

The contents of this research are structured as follows: In section 2 the background of the use of data in education to make predictions is introduced in order to give our research useful context. We explore the different machine learning algorithms that have been utilised by other researchers in related applications of prediction models in education. In section 2.3, we describe in detail the skill sets that will be used in this research and outline the significance of each of them. Section 3 provides an overview of the research methodology that will be undertaken, the preceding sections detail the undergone phases of the project and the outcome goals. Section 4 provides an overview of results and interpretations

2 Background and Literature Resources

A student's entry academic characteristics influences their success, failure or withdrawal from a particular degree [4]. Based on student pre-college academic results and demographic background attributes, researchers have made attempts to identify under-prepared students in order to initiate the facilitation of intervention programs that can help curb the attrition rate. Recommender systems based on educational data are capable of revolutionising the college experience of students for the better if they are guided into fields they are best suited for [5]. Current university entrance criteria is based on APS and key subject marks thresholds. Contrary to common misconception, meeting the admission requirements does not necessarily mean the student is suitable and equipped enough to successfully complete the degree program [6]. The high attrition rate demonstrates how ineffective the current entrance criteria is.

South Africa is undergoing transformation to address the social hindrances of it's political past, as part of the ongoing process, access to higher education has increased to cater for new groups that were previously excluded. With this transition, there has been a gap in the career guidance and support made available for many new students. As part of this research we ask to what extent the high student attrition rate is due to the misalignment of student academic skills and degree streamline? Can prospect students with limited knowledge on degree streamlines be guided into selecting career paths that they are better suited for and have a higher chance of success based on their displayed academic strengths? We propose a method to solve this issue by proposing the use of a learner skill-sets as indicators of their academic ability in order to predict their suitability for a degree. Introducing a system that can predict student suitability and prospect in a degree streamline would be beneficial to both the student and the learning institute.

The application of machine learning predictive models on university student data is capable of providing vast insights into the correlations that exist between student performance at university and their academic performances prior to university. These predictions are particularly useful in aiding students in the decision making process of selecting a degree to study towards based on their academic results pre-college.

2.1 Related Work

This research work is based on other studies that have, in different ways, investigated the link between some degree streamlines and key high school subjects capable of predicting student performance. For instance, it has been found that there is a strong relationship between high school maths results and university performance in computer science [4]. This can be attributed to the fact that computer science is based on problem solving ability which is well measured by mathematical ability.

Previous research in the prediction of student performance has found that there are certain factors that can be used as predictors of success, particularly previous academic results and biographical data [2]. We will be adding to this research field by using machine learning algorithms to assist in weighing the contributions of each of these factors when making predictions on student performance. Research in using data mining in education allows for more efficient use of resources in order to understand and predict student performances better [7].

There have been several implementations of machine learning classification models in the education space. Here we report on the findings of other researchers in the line of work closely related to this research. We compare the the approaches and factors used by the researchers in their relative studies. Background relates to authors who have used the students family attributes and demographics. These range from gender, age, schooling quantiles and their spoken languages.[8, 9]
Individual attributes include measures such as the learners interest in their studies, motivation and support structures they are exposed to. The students interaction with other peers. Schooling refers to the pre-university schooling of the student. Social system relates to the social aspects a student is exposed to such as common behaviour they are exposed to. These factors are qualitative and are not always commonly used, however some authors argue they provide more insight into the students [10].

In the related work we found that the best performing model was the SVM with an accuracy of 97%. Other researchers also used SVMs and confirm that they are capable of producing viable performances when used in used in contexts of predictions in educational data [11, 12].

Authors	Background	Individual	Schooling	Social System	Academic System	**Model Used**	**Accuracy**
Barker Kash [2004]		✓	✓			Nueral Network	61%
Nguyen Thai-Nghe [2010]			✓		✓	Logistic regression	69 %
Ajoodha et al. (2017)	✓	✓		✓	✓	Bayesian	
Lubna Mahmoud Abu Zohai [2019]	✓				✓	SVM	76.3%
Patricia F. Campbell [1984]			✓			-	-
Sonali Agarwal [2012]			✓			SVM	97.3%
Cortez Paulo [2008]	✓	✓		✓	✓	Naive Bayes	74%
Strecht Pedro et al. [2015]				✓		SVM	60%
Jaroslav Bayer et al. [2012]	✓	✓				J48	89.57
Ajoodha et al. [2019]	✓	✓	✓			Naive Bayes	69%
Mrinal Pandey et al.					✓	K-Star	85.66%

Table 2: Literature Resources - An overview of approaches used by authors in related work

In some instances, suggestions are made to adopt the integration of multiple classifiers in order to make the predictions as this offers the best model performances when predicting student performance [13]. Table 2 compares the approaches and performances of the related work investigated in line with this study.

2.2 Conceptual Framework

Figure 1 depicts conceptual framework that our structuring is based on. This an extension of a model proposed by Tinto[3]. In the original model, the relation of factors that affects a students decision to dropout are outlined. These are found to stem from the individuals own attributes that are linked to their family background and pre-college schooling. This then influences their goal and commitment which affects their grade performance and intellectual development. In our adoption of this model, we outline the fact that the individual attributed and the type and quality of pre-college schooling along with family background influences the pre-college marks of the student. These marks for this research are then each linked with a skill-set attribute. these identified skill sets then determine if the student will be able to fulfil degree requirements. Parallel to that a students characteristics influences how well they will be able to integrate with the social structure of college. All of this in turn affects the students final performance and attrition.

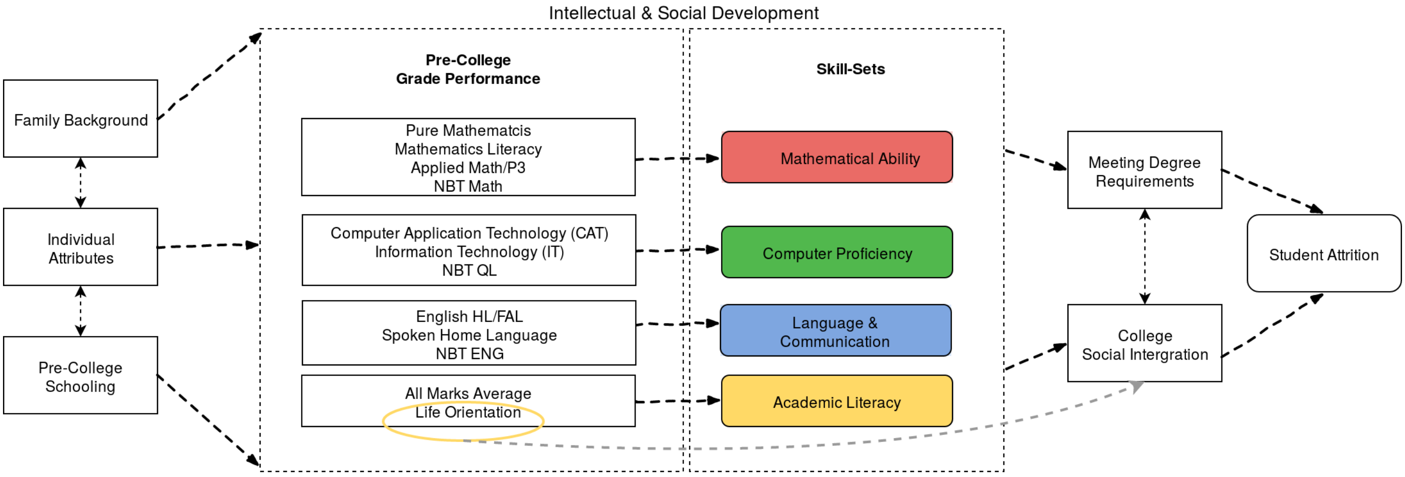


Figure 1: Applied Conceptual Framework

2.3 Skill-Sets Rationale

The readiness theory states that a prospect students readiness can be only evaluated by understanding the demands and expectations of the environment and field they are heading into [10]. The use of skill-sets is in alignment with this notion as the composition of skill-sets is based on key qualities necessary for university success. Below we describe the value and consequent justification for the use of skill-sets as utilised in this study.

Mathematical Ability

Mathematics is particularly useful in evaluating a students problem solving ability [10]. Mathematical concepts introduced in high school form the fundamentals for a science degree, thus a students performance at that level can help gauge how they might grasp to conceptual application and thinking involved in a science degree. There is strong link between mathematics and science degrees. The interpretation and understanding of mathematical concepts is also a useful measure of the students problem solving ability.

Math ability also demonstrates ones reasoning and ability to think systematically and logically. these skills are useful to measure a students lateral thinking ability which is particularly useful for finding solutions to problems in the most dynamic of environments.

Language Communication

Characterised by Language marks, number of languages studied. looking at the marks of the home language studied by the student, weighting the mark accordingly based on which level of English was studied in high school. NBT English mark The ability of the student to be able to interpret communication as well as to express themselves depends on their language capabilities and confidence. The medium of teaching at the evaluated university is English ,this means in order for a student to cope they need proficiency in English to be able to engage with lecturers as well as peers. Being able to engage in constructive and critical dialogue allows for the sharing of thoughts and ideas which allows for better learning and understanding of course material. It has been found that a students ability in the main language of instruction in education influences their academic and career progress [14].
Language proficiency is a foundational skill that influences ones ability to read which consequently is required, together with writing, for all types of learning completing assessments in education. There is a strong body of work that shows that learning problems can develop if the language in which a student has oral proficiency is not the same as the language of instruction thus as part of this research we use language as one of the skills evaluated. Essentially it represents the students displayed ability and familiarity with English and using it for academic reading and reasoning.

Computer Proficiency

Computers have become increasingly used as a medium of interaction for students and lecturers, they are more importantly a tool for obtaining further resources from the internet. One measure of how much a student has engaged with computers would be their marks in either Computer Application Technology (CAT) and Information Technology (IT). Exposure to those technical skills is useful for students given that students will primarily be engaging with computers either to access course material or the perform research and complete assessments.

Academic Literacy

Writing Ability. factored from general marks gives an overview of the students writing ability and manner in which they are able to perform generally under formal writing assessments. One powerful contributing result for this factor will be the NBT since these are assessments that one cannot necessarily prepare or study for unlike conventional schooling assessments.

The assessment methods used at university mean that one of the most important skills have to do with ones ability to write accordingly and effectively express themselves. Articulation and effectively explaining information

Academic Literacy skills are essential for success in university because of the high level of academic assessments student undergo. [15]. We incorporate life orientation in this skill-set as this has social aspects taught to learners that influences their academic ability at university.

3 Methodology

The intended purpose of this research paper is to evaluate the use of student pre-college results and home language to form skill-sets that can be used as predictors of success in Science degree streamlines. The Science degree streamlines are characterised into four major groups, namely; *Biological Sciences, Earth Sciences, Mathematical Sciences and Physical Sciences*. Learner academic ability is represented by skill-sets as follows: *Mathematical Ability, Computer Proficiency, Academic Literacy and Language Communication.*

The data used to train and evaluate the classification models was taken from a South African research intensive institute. The classification models are used to predict student degree completion by streamline based on observed skill-set attributes. The models used for this research were chosen to be exhaustive of the different machine learning paradigms, namely: Bayesian, decision trees, deep learning, instance based, ensamble and functional models. The performance of the different models under each of these domains were evaluated using confusion matrices, and the best performing models under each paradigm were selected and are reported on in this paper.

The components of this section are presented as follows: Section 3.1 describes how the data used was collected along with a comprehensive description of the data initial structuring. Further more, we describe steps undertaken to perform the pre-processing that tailored the data structure to suit our research objective. The consequent subsection discusses the composition of student skill-sets from learner pre-college attributes. This is followed by an analysis of the contribution each of them provide when used as predictors of student success. Section 3.2 outlines the generic implementation of the machine learning classification models and an overview of the evaluation metrics used.

3.1 Data Analysis and Pre-processing

In this research, data of students enrolled in the faculty of Science at a South African Research-Intensive Institute between the years 2008 and 2018 was used. This time frame was chosen because it was the time at which a new education curriculum was introduced nation wide, the National Curriculum Statement (NSC). This data, obtained from the institutes Academic Information Systems Unit (AISU), consisted of the students pre-university entrance results, demographics and within university results with final outcomes (graduation or failure/attrition).

The data received was presented as two types of table sets. The first set had student Matric Results and the Second had university course registration. In both structures, each table represented an academic year. For each year, all the undergraduate students enrolled in the faculty of science a represented. The set of course registration tables include all the courses and results the student was enrolled in for that year and their final academic outcome for the year. The matric results set of tables is linked to the previously mentioned such that it represents pre-college academic results and biographical data of all students enrolled in that particular year. In each of the tables a student is identified by an encrypted equivalent of their student number ensuring that student privacy is protected for ethical integrity.

The first step was to determine for each student which of the four streamline groups they are a part of. The science degree streamlines are characterised into four groups, namely Mathematical Science, Physical Science, Biological Science and Earth Science. This was achieved by using a probabilistic model approach based on the subjects the student was enrolled for. Certain subjects are indicative majors that influence the probability of the student being in specific science streamline. The highest probability of the four was then taken as the classification of the streamline group of the student. Students with with majors that were inconclusive in determining their degree streamline amongst the four were left out in the final data set. This was because we found that some students were enrolled under science yet have many cross faculty course enrolments making it difficult to classify their degree streamline with the Science Faculty.

The next step was to extract for each student their available pre-college (Grade 12) subjects taken and results along with NBT results where applicable since taking the NBT is not mandatory for incoming students, this component of the data was not available for all students in the data set. The single biographic feature taken was the students spoken home language. Finally for each student we determine their final qualification outcome, this is the final outcome measure of success we are interested in for the research. This is identified using the final outcome of their final year of study, or conversely the absence of a final year or other years, indicating student attrition and failure to obtain their degree.

In order to ensure that the model predictions are not biased based on the number of instances of target classes used to train and test the data, We used an under-sampling technique to balance the number of observations for each target class. We ended up with 342 instances for each class.

Column (Feature)	Type	Description
Degree_Type	Categorical {Bio; Earth; Math; Phys}	Student Science Degree Streamline
Math_Ability	Numeric 0-100	Weighted Skill-set composed of student marks
Lang_Communication	Numeric 0-100	Weighted Skill-set composed of student marks
Computer_Proficiency	Numeric 0-100	Weighted Skill-set composed of student marks
Academic_Literacy	Numeric 0-100	Weighted Skill-set composed of student marks
Degree_Outcome	Categorical {Qual; Qual_Late; Fail}	Degree successful completion (Record Time or Late) or failure

Table 3: Breakdown of extracted student table csv

Composition of Skill-sets

Based on the outlined properties of each of the skill-sets mentioned in section 2.3, we classify student marks into respective skill-sets to profile the students ability. Table 4 outlines the composition of skill-set in term of the observed subject marks below. Each mark is associated with a weight based on the subjects impact relative difficulty in comparison to other similar subjects. Due to the variation in grade 12 subject names we found in the data structure, it was necessary to identify subject codes comprehensively.

The mathematical ability of a student is primarily based on the students Pure Mathematics mark, students who alternatively took math literacy have the contribution of that mark at a much lower weighting. Additional math subjects are weighted slightly above the NBT Math as their presence shows an additional exposure to mathematical

concepts. $0 \leq$ Skill-Set Value≤ 100
Math ability: Pure Mathematics (*MATNSC, MAT, MAT-HG*),
Math Literacy *MALNSC,MAT-SG*, Mathematics Paper 3 (*MP3NSC*), Applied Mathematics (*APMNSC, ADV, ADM-HG*) and NBT Math (*NBTMA*).

The language communication ability places emphasis on English as this is the medium of communication at South African university institutes therefore it takes precedence. An additional biographical data attribute is included which is the mother tongue or spoken home language of the student. The NBT AL, Academic Literacy is included here and not as part of the academic literacy skill-set due to fact that upon close investigation the structure of the assessment actually evaluates a students ability to read and interpret instruction, that level of comprehension involved is a measure of the students English language interpretation and understanding which consistent to how university tests and examinations would be structure requiring a similar level of insight and interpretation. **Language Communication:** English Home Language (*ENANSC, ENA-HG, ENH*), English First Additional Language (*ENBNSC, ENL, ENB-HG,ENCNSC*). NBT Englsh (*NBTAL*).

Familiarity to computers influences the experience the student may have engaging with technology during their time at university especially since online platforms are more widely used for resource access and the submission of material. Information Technology is a lot more technical than Computation Applied Technology therefore it has been given a slightly higher weighting. The NBT QL (Quantitative Literacy) assess the students ability to solve problems in textual, tabulated and graphical environments. **Computer Proficiency**: Information Technology (INTNSC), Computer Applied Technology (*CATNSC, CSC-HG, CST, CST-HG,CST-SG,CSC-SG*) NBT Quantitative Literacy (*NBTQL*).

As part of the academic literacy skill-set we incorporate the students all other marks, this is to provide an overview of the student general academic ability. An incorporation of the life orientation mark is due to the importance of life skills as part of social skills that influence the ability of the student to adjust to university... Should be called life skills and academic ability. **Academic Literacy**: Life Orientation (*LFONSC*), All leaner matric subjects average.

Table 4 below shows the established weighting of each of the contributing marks for the four respective skill-sets.

Skill-set	Subject	Weighting
Math	Pure Math	0.5
	Math Literacy	0.25
	Math P3 or AP Math	0.3
	NBT MAT	0.2
Computer	Computer Application Technology	0.65
	Information Technology	0.7
	NBT QL	0.3
Communication	English HL	0.5
	English FAL	0.3
	NBT ENG	0.3
	English Spoken Hl	0.2
Academic Ability	Life Orientation	0.5
	All marks Avg	0.5

Table 4: Skill-set mark composition weightings

3.2 Applying Classification Models

The Machine Learning models were taken to exhaustively cover the paradigms of machine learning. In each of the classification instances 10 fold cross-validation was used. A total of 1026 student records were extracted and used by the models. Initially the class representations were imbalanced, to ensure that the accuracy performance of the classification models is not biased, we used under-sampling to balance the classes ensuring that there was an equal representation of each one. In each class of qualified, late qualified and fail there were 342 instances, thus making it a total 1026 records used for the implementation of the approaches described below. The six generic models used are listed below :

Bayesian:

Bayesian Network - This is a probabilistic model that uses Bayesian inference to calculate the probability of a degree of causation. Using prior probability of instances feature A given class B. The Bayes rule can be interpreted as: $P(Y|X) = P(Y)\frac{P(X|Y)}{P(X)}$ where $P(Y|X)$ represents the desired posterior probability. $P(X|Y)$ represents the prior probability.$P(X)$ and $P(Y)$ are the features observed and target class respectively [16]. Figure 2 depicts the structure of the Bayesian network for this work, it shows the fact that the final outcome is dependent on all of the attributes it point towards below.

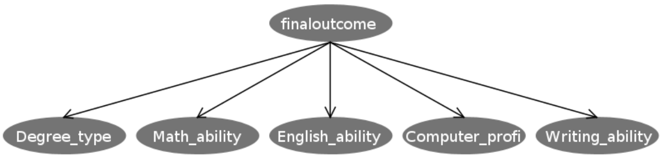


Figure 2: An illustration of Bayesian Network

Decision Trees:

Random Forrest - This method works by dividing the data into two disjoint subsets, one that is operated on by the decision tree, the other by the naive Bayes. A technique called forward selection search is used within the steps where attributes are modelled by naive Bayes and the other remaining ones by the decision tree. Random forest, as the name suggests, is composed of numerous decision trees that operate as an ensamble which adopts a combination of results and implementation of different models and instances. These uncorrelated decision trees have their properties combined and work in conjunction in what is referred to as a committee, this then outperforms any on the individual models of trees. [17]

Deep Neural Learning:

Multilayer Perceptron - A feed forward neural network that uses back propagation to learn the parameters. This model is a neural network that comprises of multiple layers and at each of the layers the nodes are sigmoid functions. In instances where the class is numeric the output nodes become linear units which are unthresholded [18].

Ensamble (Meta) -

Bagging (Bootstrap Aggregation)- This model incorporates many methods of approach into one. It makes use of statistical classification and regression. The mode prioritises the reduction in variance, noise and bias. This is particularly due to the bootstraping aspect of this model which uses random sampling procedure with replacement. With each sample set taken, decision trees are used to learn the classification. At these stages there is no concern of overfitting and the trees can run deep. When all the decision trees are learnt they are then combined accordingly to form the one model used for prediction [19]

Functional:

Support Vector Machine - This works by separating classes of the training data by using a multi-dimensional hyperplane. The predictions made on the test data instances work by mapping each instance into a side of the hyperplane which represents the predicted class. Generally the paradigm of functional machine learning models work by fitting functions to curves. In this study we used the fine Gaussian SVM which means it works by making finely detailed distinctions between classes with the use of a guassian kernel.

Instance (Distance) Based:

K-Star - This model is based on how close instances are to each other and then analysis is made on this. This works by calculating the distance using an entropy-based function to classify test data instances based on training data formulations [20].

All the above mentioned classification models will be evaluated using a confusion matrix and consequently the accuracy of the model

3.3 Ethics Clearance

The permission for the data used from a South African Research Intensive Institute, was obtained through the human research ethics committee, under clearance certificate protocol number H19/09/24.

4 Results and Discussion

This section presents the results of the classification models used in this research. We train and tested the models to predict the university outcome of a student based on their skill-set attributes for respective degree streamlines. In subsection 4.1 the performances of these models are presented using confusion matrices. Subsection 4.2 provides analysis of the performance of the models with the use of graphical Interpretations particularly ROC Curves and a standard bar graph for an illustrated comparison of the accuracy of models. subsection 4.3 Provides an analysis of the information gain provided by each of the skill-sets when making classifications.

4.1 Classification Models Predictions

Below we present the results of the six selected classification models, Namely Bagging, Bayesian Network, K-Star, Multilayer Perceptron, Random Forest and SVM. Figure 4 illustrates the results of these model algorithms.

Figure 4a illustrates the confusion matrix of the *Bagging* classification model which achieved an accuracy of 73.5% using 10-fold cross validation. Notable misclassifications observed saw the confusion between students that Qualified in record time (Represented by Qual) and those that Qualified taking more time then the minimum required years. The missclassification of Qual as Late Qual is at 24.%. Closely behind that, late qual was misclassified as Qual 23% of the instances. This model took 0.16 seconds to build.

Figure 4b illustrates the confusion matrix of the *Multilayer Perceptron* classification model which achieved an accuracy of 68.5% using 10-fold cross validation. This model has a significantly higher missclassification of Qual as Late Qual at 29% and conversely 33.% of Late Qual as Qual. This is model has the highest level of misclassified instances of Fail and Late Qual at 12% similarly so in the converse. This model took the longest of time to build compared to all other models taking 1.09 seconds, followed by the random forest model, all other models built in time that is approximately 0 seconds

Figure 4c illustrates the confusion matrix of the *K-Star* classification model which achieved an accuracy of 80.4% using 10-fold cross validation. This model has the joint highest number of correct classified instances of fail at 89% together with the Random Forest model. This model has the highest correct classifications of late qual at 89% which the most misclassified class in other models. Overall this model has the highest accuracy performance compared to all others used in this research.

Figure 4d illustrates the confusion matrix of the *SVM* classification model which achieved an accuracy of 77.4% using 10-fold cross validation. This model has the highest correct classification of Qual at 78%. The missclassification of late qual as qual is at 22.8%

Figure 4e illustrates the confusion matrix of the *Bayesian Network* classification model which achieved an accuracy of 65.5% using 10-fold cross validation.This model also has significantly high misclassifications of Late qua as qual at 38.3% with Qual misclassified as late qual 33.6% of the instances. This model has the highest missclassification of Fail as late qual at 16%. This model took 0.04 seconds to build.

Figure 4f illustrates the confusion matrix of the *Random Forrest* classification model which achieved an accuracy of 77% using 10-fold cross validation. This model has the joint highest correct classification of fail shared with the k-star model. This model also the most missclassification confusion happening between qual and late qual, qual is misclassified as late qual 22.5% and conversely the opposite 21.9% . This model took the second longest amount of time to build, taking 0.68 seconds.

The K-Star classification model achieved the best accuracy performance, after investigating the underlying functioning of this mode, the performance can be attributed to the fact that this model uses an entropy-based distance function. One of the benefits of this approach is that it handles missing values well [20]. In our data set, the skill-set of computer proficiency had many missing values due to the fact that not all students take a computer related subject in high school. This aspect is seen clearly in figure 8, the box plot showing the distribution of skill-set valuations accross the different streamlines show that there are significantly lower marks because of their absent. Similarly so, the NBT is not a mandatory assessment at the institution that this research is based on. Therefore these missing values are likely to have influenced the other models relatively while the K-star was able incorporate them accordingly when learning and making predictions.

Presented below are the confusion matrices of the classification models. Figure 3 gives a breakdown of how the results of the confusion matrices are interpreted.

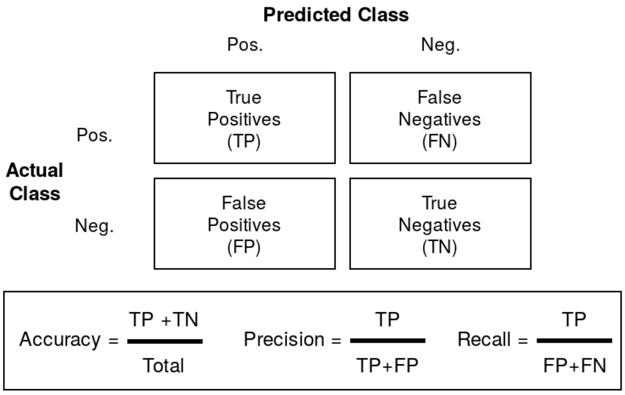


Figure 3: Confusion Matrix breakdown

		Predicted		
		Qual	Late Qual	Fail
Actual	Qual	245	85	12
	Late Qual	81	225	36
	Fail	20	38	284

(a) **Bagging** achieves **73.5 %** accuracy. with 754 correctly classified instances and 272 incorrectly classified instances.

		Predicted		
		Qual	Late Qual	Fail
Actual	Qual	234	100	8
	Late Qual	113	187	42
	Fail	18	42	282

(b) **Multilayer Perceptron** achieves **68.5 %** accuracy. with 703 correctly classified instances and 323 incorrectly classified instances.

		Predicted		
		Qual	Late Qual	Fail
Actual	Qual	268	66	6
	Late Qual	65	252	25
	Fail	18	19	305

(c) **K-Star** achieves **80.4 %** accuracy. with 825 correctly classified instances and 201 incorrectly classified instances.

		Predicted		
		Qual	Late Qual	Fail
Actual	Qual	277	57	8
	Late Qual	78	233	31
	Fail	8	25	284

(d) **SVM** achieves **77.4 %** accuracy. with 794 correctly classified instances and 232 incorrectly classified instances.

		Predicted		
		Qual	Late Qual	Fail
Actual	Qual	215	115	12
	Late Qual	131	182	29
	Fail	12	55	275

(e) **Bayesian Network** achieves **65.5 %** accuracy. with 675 correctly classified instances and 351 incorrectly classified instances.

		Predicted		
		Qual	Late Qual	Fail
Actual	Qual	253	77	12
	Late Qual	75	232	35
	Fail	14	23	305

(f) **Random Forest** achieves **77 %** accuracy. with 794 correctly classified instances and 232 incorrectly classified instances.

Figure 4: Confusion matrices describing performances of classification models.

Model	Accuracy	Precision	Recall	F_Measure
KStar	80.4094	0.805	0.805	0.804
SVM	77.4131	0.774	0.775	0.774
Random Forest	76.9981	0.768	0.77	0.769
Bagging	73.4893	0.737	0.735	0.736
Decision Table	70.2729	0.706	0.703	0.704
Multilayer Perceptron	68.5185	0.686	0.685	0.685
Bayesian Network	65.4971	0.663	0.655	0.658

Table 5: Performances of the classification models

4.2 Graphical Interpretations

Below we present graphical representations of the performance of the classification models, Figure 5 depicts the results tabulated in table 5. This was done to compare in more detail the performances of the classification models, In addition to the accuracy, we also evaluate the precision and recall. How these were calculated from the confusion matrices is given in detail in figure 3. We found that precision and recall were very close to the accuracy and thus gave arguably the very closely similar interpretations.

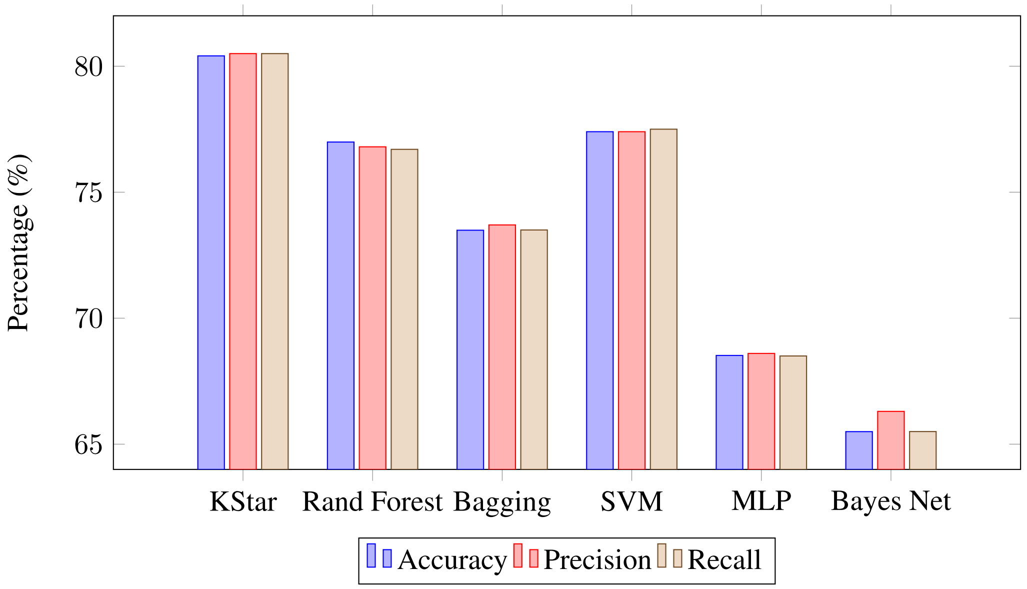


Figure 5: Bar graph illustrating classification model performances

ROC (Threshold)Curve

Here we describe the results of the Receiver Operating Characteristic (ROC) and the subsequent Area Under Curve (AUC) below the plotted lines. This approach sees the use of cost sensitive classifiers, thus it was important that our classes were balanced to ensure the results were not bias and thus credible.Figure 6 depicts the ROC curve that shows the true positive rate versus false positive rate for selected target output class at various thresholds, also known as threshold curve. In this depicted instance, analysis on the best performing classification model, K-Star is evaluated. The 3 target classes, Qualified, Late Qualification and Failure classes are illustrated.

The Area Under Curve (AUC) values of each of the classification models is depicted in table 6 with the average of the entire model shown. The values of this section range from 0 to 1 where values greater than 0.5 are considered to be good since this means it performs better than random. In our case all our models achieved satisfactory performances

We found that the rankings of the models based on their average AUC differs from the ranking of the models based on their accuracy. For instance when ranking the models in descending order of accuracy, SVM ranked second. However in the AUC rankings the random forest model was ranked second, followed by SVM. The rest of the models maintained the rankings found when evaluating the accuracy.

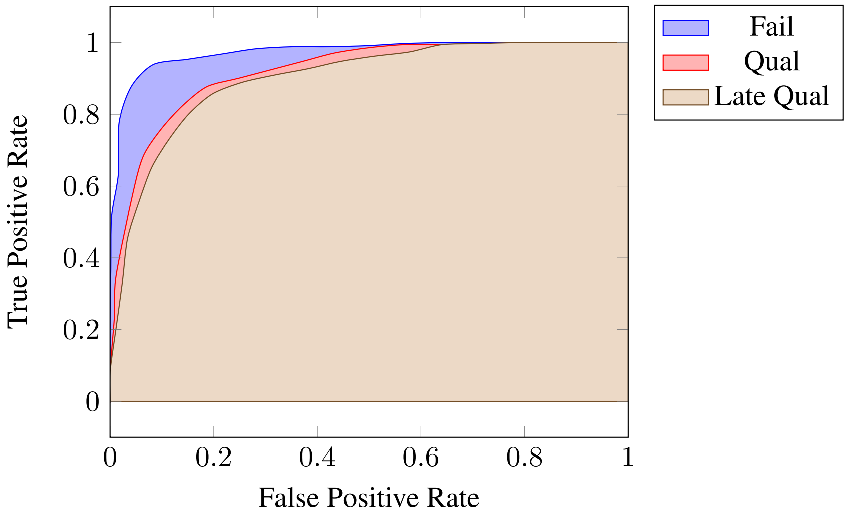


Figure 6: ROC Curve of KStar Model

Model	**Area Under Curve (AUC)**			
	Qualifed	*Late Qualification*	*Fail*	**Average**
K-Star	0.9210	0.8985	0.9732	0.9309
Random Forest	0.9050	0.8737	0.9642	0.9143
SVM	0.9012	0.8401	0.9632	0.9015
Bagging	0.8774	0.8387	0.9544	0.8896
Decision Table	0.8422	0.7741	0.9424	0.8529
Multilayer Perceptron	0.8439	0.7611	0.9342	0.8464
Bayesian Network	0.8221	0.7501	0.9418	0.8380

Table 6: AUC of classification models

4.3 Information Gain Attribute Evaluation

In this subsection we compare the entropy rankings of each of the skill-sets in order to establish how well each of them contribute to the classification of student final outcomes. This is also to investigate how much influence a skill-set has in the final university outcome prediction of a student. The rankings and results are tabulated in table 7 and depicted as a line graph in figure 7.

With this investigation We found that learner writing ability significantly has the highest information gain, followed by Mathematical ability. Language and Computer Proficiency were the least ranked so much so that they could be interpreted as interchangeable.

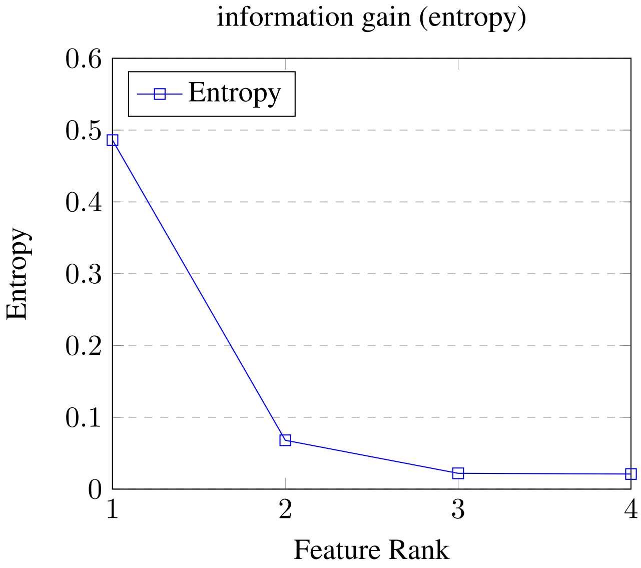


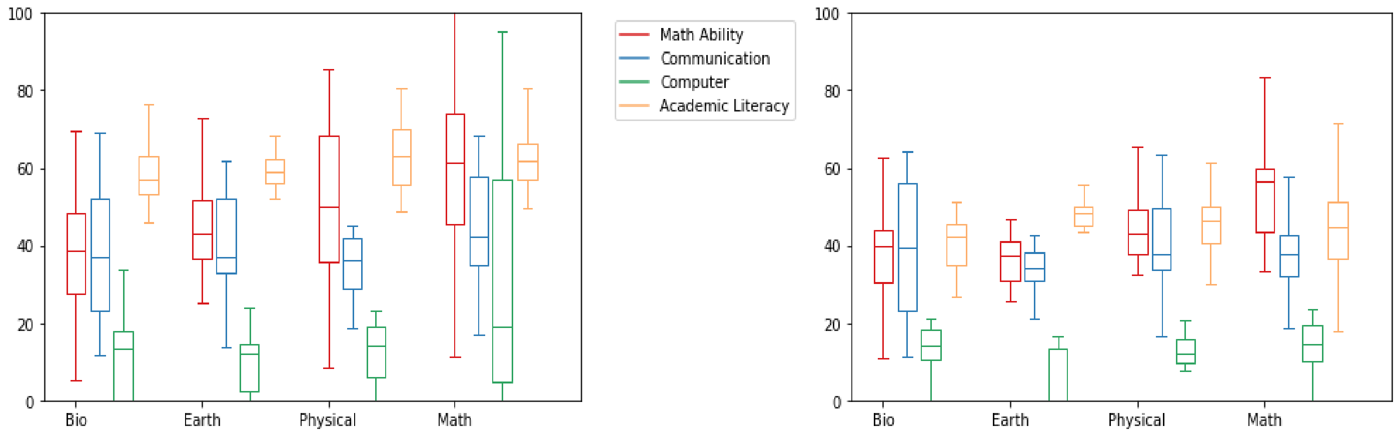
Figure 7: Line graph plot of entropy of features

Rank	Entropy (e)	Feature
1	0.486	Writing Ability
2	0.068	Math Ability
3	0.022	Language Communication
4	0.021	Computer Proficiency

Table 7: entropy ranking of features

4.4 Skill-set attributes across streamlines

In this subsection we investigate how the skill-set values are distributed and differ accross the different degree streamlines. Figure 8 depicts the box and whisker plots of students a) who have qualified and b) students who have failed to qualify and obtain their degree.



(a) Qualified Students

(b) Failed Students

Figure 8: Box and Whisker Plot of Skill-set distribution by Degree Streamline

Through the investigation of these box plots we found that students who failed have Maths ability medians that are relatively close. Contrary to initial assumption made before the inception of this study, maths ability does not immediately distinguish student who will fail or pass. When looking at language and communication as a skill, failed Students in Physical Sciences have higher communication then qualified students, this can be interpreted it meaning the students who failed likely speak more languages than those that qualified.

5 Conclusion

In conclusion of this study we reiterate the fact that tertiary education is marred with a lot of uncertainty when it comes to the prospects of new students, many fall along the way struggling to keep up with the demands and pressures of the tertiary learning environment, with many failing to obtain their degree qualifications. With this model approach investigated in this research paper we attest that it is a viable alternative to the conventional APS entrance criteria. In this work we used learner skill-sets to represent their academic ability when entering university. We introduce this configuration in contestation of the conventional APS model in order to evaluate a students suitability for a science degree streamline.

The purpose of the model adopted is to better represent a students abilities. This approach gives a more holistic view of what the student is capable of in the academic space. Skill-sets are a means representing characteristics with the advantage of being able to identify early on any aspects the student may be lacking in for possible intervention and support unlike the APS approach which gives no such insights. The contributions of this body of work include a comprehensive evaluation and comparison of the best performing machine learning classification models. These algorithms identified are capable of predicting student suitability for a degree streamline providing a means to for-see and prevent student attrition. The best performing classification model achieved an accuracy of 80.4% which is in close proximity of some of the related work reviewed.

Further contributions of this study include the analysis of skill-sets in terms how they each contribute to the prediction process of student outcome. We evaluated the information gain, also referred to as entropy, of each of the skill-sets. After finding the entropy values of each of the skill-sets as feature we then ranked them accordingly. We found that a students writing ability had the highest entropy followed by math ability. The other two skill-sets had entropy values that was significantly lower yet close to each other, thus we concluded that the last two of language communication and computer proficiency can be interpreted as being interchangeable.

The last major contribution of this paper is to have the best performing classification model, K-Star, as the back end functioning of a proposed application. This application is aimed at prospect university students intending to pursue a career in the science faculty. Subsection 5.1 explains the details of this application further with the consequent application prototype shown in figure 9.

limitations

As with any research approach, this study too has its limitations. Firstly during the pre-processing and data cleaning stages, many student records had to be removed as they has attributes that were incompatible with the purpose of this study, firstly in many cases we were unable to classify many student degree streamlines into one of the four. This was due to the fact that some students had university courses that they were registered for that stretched into different faculties and in some instances overlapped the four degree streamlines. The probabilistic approach used to classify the degree streamline also in some instance gave balanced probabilities that were inconclusive. Other student records that were removed were students that had no marks in the key subjects that are mandatory for all students in some way, such as having no maths or English related subject. This could have occurred because of human error at some point in capturing the data.

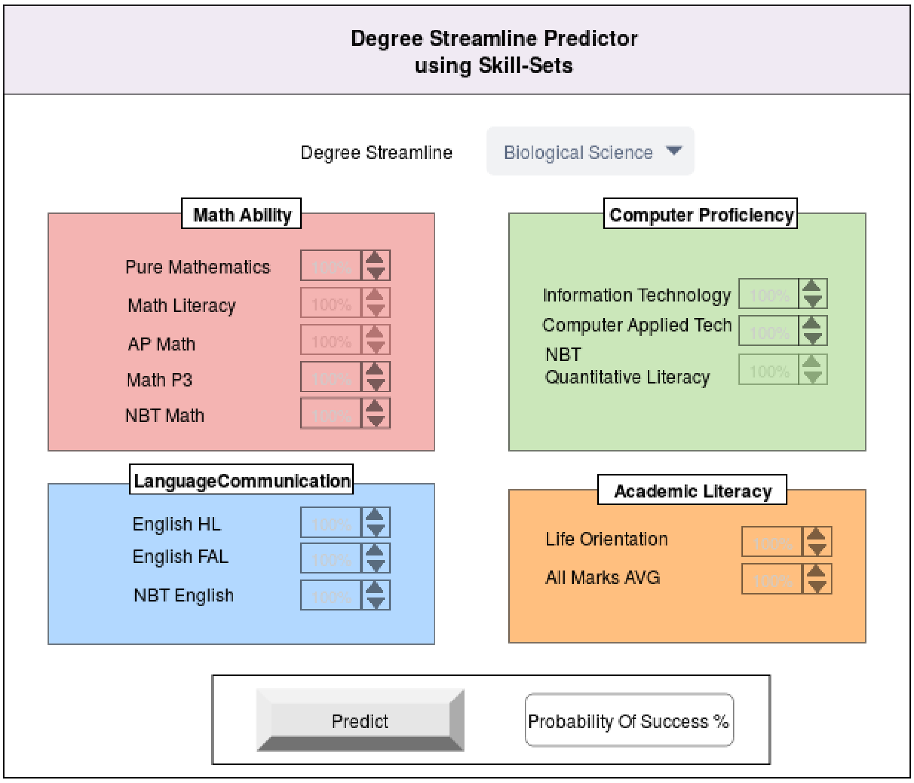
Another limitation of this is that skill-sets are only composed of quantitative data, there are other student attributes that are important predictors of student attrition as described in the Tinto conceptual framework [3]. For instance factors such as a students self-efficacy , motivation and goal commitment currently has no means to be measured in the academic space. In the language communication skill-set a measure of the students oral measure of expression is limited. This is a key skill for engaging in lectures, consultations and presentations in the university space which also influences a students probability of success.

5.1 Future Work

The work of this research can be extended further by other researchers in order to represent a learners characteristics even more holistically than the implementation of this study. The incorporation of the students background and demographics gives better context of their pre-university marks. For instance, evaluating the quartile of the students high school gives a better representation of the students quality of education in high school as well as some measure of the facilities and resources they had at their disposal which influences their high school academic performances.

The use of psychometric evaluations can be used to better understand a students interests, career goals and other insights such as the students behavioural style [21] Future Work-Social Integration, demographics, motivation,ambition (Tinto). Psychometric evaluations.

Below we see the proposal of an application that can be used to perform the prediction of a new student given their academic scores logical presented as skill sets. This prototype can be extended into an application that can be used by prospect students to find out their probability of success in different science degree streamlines.



Degree Streamline Biological Science

Figure 9: App Prototype

Concluding Remark...For prospect university students, writing ability. When the research began, it was expected that maths ability would be a major indicator of success. After evaluating the data our understanding has been changed to know that remediating writing ability is what we need. Not necessarily maths tuition, students being able to express themselves. Academic Literacy problem. Academic discourse is far more complex and it requires more attention

References

[1] I. Scott, N. Yeld, and J. Hendry, "A case for improving teaching and learning in south african higher education," *Higher education monitor*, vol. 6, no. 2, pp. 1–8, 2007.

[2] R. Ajoodha and A. Jadhav, "Identifying at-risk undergraduate students using biographical and enrollment observations for mathematical science degrees at a south african university," *International Journal of Science and Mathematics Education*, pp. 1–21, 2017.

[3] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of educational research*, vol. 45, no. 1, pp. 89–125, 1975.

[4] P. F. Campbell and G. P. McCabe, "Predicting the success of freshmen in a computer science major," *Communications of the ACM*, vol. 27, no. 11, pp. 1108–1113, 1984.

[5] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, "Recommender system for predicting student performance," *Procedia Computer Science*, vol. 1, no. 2, pp. 2811–2819, 2010.

[6] M. Wilson-Strydom, "Multi-dimensional approach to readiness for university," *Senior research fellow in the Centre for Research on Higher Education and Development at the University of the Free State.*, pp. 1–2, 2015.

[7] S. Agarwal, G. Pandey, and M. Tiwari, "Data mining in education: data classification and decision tree approach," *International Journal of e-Education, e-Business, e-Management and e-Learning*, vol. 2, no. 2, p. 140, 2012.

[8] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[9] J. Bayer, H. Bydzovská, J. Géryk, T. Obsivac, and L. Popelinsky, "Predicting drop-out from social behaviour of students.," *International Educational Data Mining Society*, 2012.

[10] D. T. Conley, "Redefining college readiness.," *Educational Policy Improvement Center (NJ1)*, 2007.

[11] L. M. A. Zohair, "Prediction of studentâs performance by modelling small dataset size," *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, p. 27, 2019.

[12] K. Barker, T. Trafalis, and T. R. Rhoads, “Learning from student data,” in *Proceedings of the 2004 IEEE Systems and Information Engineering Design Symposium, 2004.*, pp. 79–86, April 2004.

[13] M. Pandey and S. Taruna, “Towards the integration of multiple classifier pertaining to the student’s performance prediction,” *Perspectives in Science*, vol. 8, pp. 364–366, 2016.

[14] R. Barac and E. Bialystok, “Bilingual effects on cognitive and linguistic development: Role of language, cultural background, and education,” *Child development*, vol. 83, no. 2, pp. 413–422, 2012.

[15] D. Andrews and R. Osman, “Redress for academic success: possible’lessons’ for university support programmes from a high school literacy and learning intervention: part 2,” *South African Journal of Higher Education*, vol. 29, no. 1, pp. 354–372, 2015.

[16] D. Koller, N. Friedman, and F. Bach, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[17] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] T. Mitchell, B. Buchanan, G. DeJong, T. Dietterich, P. Rosenbloom, and A. Waibel, “Machine learning,” *Annual review of computer science*, vol. 4, no. 1, pp. 417–433, 1990.

[19] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[20] J. G. Cleary and L. E. Trigg, “K*: An instance-based learner using an entropic distance measure,” in *12th International Conference on Machine Learning*, pp. 108–114, 1995.

[21] S. Hammond, “Using psychometric tests,” *Research methods in psychology*, vol. 3, pp. 182–209, 2006. Available at `books.google.com`.

Declaration
University of the Witwatersrand, Johannesburg
School of Computer Science and Applied Mathematics
Senate Plagiarism Policy

I, INSERT YOUR NAME HERE, (Student number: INSERT YOUR STUDENT NUMBER HERE) am a student registered for INSERT THE COURSE CODE HERE in the year INSERT THE ACADEMIC YEAR HERE.

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 ALL the work submitted for assessment for the above course 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 may take disciplinary action against me if there is a belief that this in not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature:

Signed on 21 day of October, 2019 in Johannesburg.