Data-driven Student Support for Academic Success by Developing Student Skill Profiles

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

Shalini Dukhan
School of Animal, Plant
and Environmental Sciences
The University of the Witwatersrand
Johannesburg, South Africa
shalini.dukhan2@wits.ac.za

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

***Abstract*—In this paper, we attempt to provide a data-driven solution to the data-congested environment of attributes related to student success and contribute towards preventing the increased dropout rates at South African higher education institutions.**

One of the most significant discussions in higher education is student attrition in their first year of study. Student career guidance is an area that requires investigation in light of high attrition rates at university. Recent developments in data analytics, and the analysis of large data sets have enabled the production of powerful predictive models. This paper highlights how a predictive model can assist students, with an interest in Science to develop a skill profile required to be successful in their undergraduate Science programme. This is achieved by identifying the difference between the necessary skills required to be successful in a science programme (derived using data driven approaches) from the current learner's skill profile (derived from the learners' performance in assessments). The learners' skill profile is used to predict success in four Science streamlines. Based on the prediction results, we gauge the improvement in skills required to succeed in that programme.

We provide the following contributions: (a) a trained classifier able to calculate the distribution over learners' success in Science streamlines focused around the notion of skill profiles; (b) a ranking of these skill profiles according to their information gain (entropy); and (c) an interactive program to calculate the posterior probability over these skill profiles given learner's pre-university observations. We argue that it is crucial that students gauge the focus areas of skill improvement prior to enrolling for their degree so that they can consider streams in Science degrees that are suited to their academic strengths.

***Index Terms*—Suitability of Science streamline, Schooling Observations, Machine Learning Classification, Student Attrition**

I. Introduction

While Big Data has informed decisions mainly in business, government and healthcare, the problems faced by higher education institutions could also be addressed by Big Data Analytics. Students who enrol into the Sciences at University may not always be aware of the skills necessary to succeed in their chosen undergraduate programme. These skill-sets can be broadly classified as language skills; mathematical skills, computer skills, and life orientation skills.

It is necessary to determine the extent to which the skill-sets specifically impact the student's success within degrees in the Science streamlines. This paper addresses this need. This is achieved by identifying the difference between the necessary skills required to be successful in a science programme from the current learner's skill profile. The learners skill profile is then used to predict success in four Science streamlines. Based on the prediction results, we can gauge the improvement in skills required by the learner to succeed in that programme.

Following the conceptual framework of [1] we define several features associated with background, individual attributes, and high-school grades which can be used to build a skill profile for a learner to classify into one of four Science streamlines to assess the degree of skill improvement required.

We trained several machine learning classification models such as decision trees; instance-based classifiers; naïve Bayes models; neural networks; support vector machines; and linear logistic regression models to deduce the learner's skill profiles into these Science streamlines. Confusion matrices were used to gauge model performance and factor analysis was performed to rate the information gain of each feature to predicting the class label.

The results indicate that the learner's mathematical skill contributed the most towards classifying the student into a Science streamline followed by computer skill; language skill; and lastly life orientation.

The best reported accuracy was the multilayer Perceptron which achieved 62% over the four classes. A WebApp has also been prepared which uses the multilayer Perceptron classification model to categorise a learner into these four Science streamlines using Pre-college and Schooling as input features as well as to access the improvement of the skill profile necessary to succeed in each streamline. The WebApp is currently available at http://matlabapps.ms.wits.ac.za:9980/webapps/home.

The main contribution of this work is from a learner-centered perspective. We provide a data-driven approach to quantify the skill-set improvement necessary for a learner to succeed in a Science streamline. More specifically, we provide the following contributions: (a) the first trained classifier able to calculate the probability of a learner's assignment to a Science streamline around the notion of a skill profile; (b) a ranking of these skill profiles according to their information gain (entropy); (c) an interactive program which is able to calculate the posterior probability over these skill profiles so that

the improvement of the skill-profile can be measured for the learner to succeed in the undergraduate Science curriculum.

We note that the Multilayer Perceptron (a feed-forward neural network) achieves the best performance with a 62% accuracy over the class variable, which is significantly better than a random assignment.

This document is structured as follows. Section II highlights the state-of-the-art contributions in the domain of predicting a Science Streamline and a selected conceptual framework for student attrition with respect to learner skill profiles; Section III highlights our data, feature selection, and choice of classification models; Section IV outlines our major findings; and Section V concludes this paper, outlines our contributions, and puts forward recommendations for future work.

II. Literature Review

Globally, at universities and especially within the Sciences, there has been a move for providing support programmes for under-prepared undergraduates. Issues of challenges experienced by undergraduate students have been met with arguments for the academic literacies that they have to be offered within the academic context [2]. This is especially true when trying to remove the lens of the deficit model and trying to understand student attrition and transition. But, the question that begs an answer is to what extent are those students, who are enrolled in support programmes, there due to a mismatch between their skills-set and the expectation of skills sets required to succeed? Can we predict the skill-set improvement necessary - for a particular Science streamline - that the learner needs to achieve to succeed in that streamline?

This paper provides an answer to this question by offering an application which can, on the basis of a set of academic skills, propose the most appropriate skill-set adjustment for the learner to undergo. This could offer the student a mechanism for support prior to entering into university, and this study is especially important due to the limited career guidance and academic counselling available for students from all walks of life and at all levels of study. It is possible that if students are schooled to make more informed decisions on their academic trajectories and career pathways and thus on their options for studying towards a professional qualification in Science, they can increase their chances of academic success. This could therefore, have a positive effect on the rising attrition rates at universities. An application that has been developed as part of this study is proposed as one possible mechanism which can be used to guide students who would like to study within the Sciences. The application is also able to provide guidance on the most appropriate streamline in Science for the student based on their skill strengths.

Within developing economies such as South Africa, and with the need for transformation and social redress, it is essential to give students the best possible chance for academic success. High rates of attrition have negative consequences at various levels such as at institutional level- where there would be loss of finances due to investments made in educating students who ultimately drop-out, teaching staff who have invested in the teaching and learning environment, bursars of students, and in terms of the students, the sense of self-efficacy. The students' sense of academic self-efficacy is a significant factor to consider when deliberating on methods to improve academic performance [3]. Students who are academically well-suited for a particular streamline are likely to be motivated to apply themselves in this area and thus more likely to achieve academic success. It is therefore necessary to provide students with guidance on the streamline that is best suited to their academic attributes.

The level of preparation in mathematics at secondary school, and the grade achieved within this context, is shown to be closely related to students' academic achievement in the Sciences at university (see for example [4], [5]). Thus, within the context of the application developed in this study, it is considered important to include different levels of mathematics which can be taken at high school. These levels include core mathematics, mathematics literacy, and the national benchmark test for mathematics that is taken as an entrance assessment for some universities in South Africa.

Additionally, within the area of computer sciences there is also a relationship between the student's intentions to complete their studies and their degree completion. [6], this speaks directly to the level of guidance that students receive when making a choice for their studies and professional career paths. An application, such as the one proposed in this paper, could assist students to better understand the streamline that is preferable for them based on their academic strengths and skills and the need to develop them. The application thus accounts for the inclusion of computer science grades and the national benchmark test for quantitative literacy as well. These two factors broaden the spectrum of variables that the model can use to more powerfully predict the suitability of the skill-set for a streamline.

Language has been shown as closely linked to academic performance in Science for undergraduates (for example [7]–[9]). Thus, in this research paper we take into consideration the school-leaving language score as well as the mother tongue of the student. Life orientation scores have additionally been included since this type of subject is meant to assist students with positively meeting the demands at the social level, enables self-regulatory and self-responsibility capacities, as well as gears them to optimise their chance of success in adulthood [10].

Machine learning algorithms have been used in this study to differentiate between the predictive power of different classification models. The classification models enable the interpretation of the inter-relationship between different factors included in the analysis. Thus, it is possible to compare the predictive power across a variety of models [11]. Little is known about how the predictive models, based on machine learning algorithms, could facilitate and guide the students' selection of a streamline within the Sciences at university. This study sheds light on this area, and would be useful to university student support programmes and career guidance centres. The application that has been developed can also be

used by the student as an individual.

Statistically, the success of a student in a degree/course is correlated with their performance in previous assessments (i.e. grades in mathematics and computer science, and home language) [12]. Life orientation is also important in terms of providing the student with the skills to succeed in the environment.

Students could benefit in better understanding which streamline in Science is best suited to their academic aptitude from the application proposed in this study. For the university, optimising student success means more skilled graduates, lower-dropouts, higher pass-rates and increased employability. Finally for the lecturer, since student gaps in skill will be continually focused on and developed, the lecturer can cover more specialised content. Moreover, this justifies the efforts of lecturers to improve their teaching, thus impacting academic development initiatives positively.

[13] and [14] summarised the influencing predictors into multi-dimensional factors, namely prior to admission student centric (age, ethnicity, prior academic performance, gender, time management, information / computer literacy, reading and writing capability), after admission academic centric (academic integration, clarity of programs, interpersonal relationship, study habits, accessibility to services, absenteeism, social commitment etc.,) and external environment (financial status, family responsibilities, external support, life crisis etc.,) influencing students decisions.

In the last $\approx$ 25 years the analysis of student data has transformed from mining of survey data [15], use of principal component analysis by [16] to use of machine learning to predict student dropouts by [17]. In [18] of 450 students who enrolled in 71150 information management courses over three years (2006-2009), showed that socio-demographic predictors like ethnicity, course programme and course were significant predictors of students being successful or not. He also showed that risk estimated based on enrolment data is not enough to predict the student fate. [19] found for courses in mathematics and computing, the biggest factors for student attrition were the marks in first assignment, course level followed by course rating, course work, gender, age and socio-economic status of the students.

[20] compared various supervised machine learning algorithms like decision trees, artificial neural networks, naïve Bayes classifier, instance-based learning, logistic regression and support vector machines to predict student's success and showed that when only demographic predictors are used, the accuracy was 58.84%, whereas when other predictors were used along with demographic predictor, naïve Bayes classifier accurately predicted performance of the student.

III. Methodology

In this paper we attempt to profile learners by their skill-set in order to identify the necessary skill improvement that can result in the success of the learner in a science streamline. Towards this goal, we attempt to predict the success of skill-sets in four science streamlines. We will train several machine learning classification models from different archetypes of machine learning to deduce the suitability of a learner's skill-set in these four Science streamlines. Confusion matrices will be used to gauge model performance and factor analysis will be performed to rate the contribution of each feature to predicting the class label.

A. Data Collection and Pre-processing

The data consisted of enrolment observations of students from the Faculty of Science degrees at a South African university. These degrees included streamlines of specialities from four major science streamlines: Earth Sciences, Biological and Life Sciences, Physical Sciences, and Mathematical Sciences. The enrolment observations of each learner was collected for learners registered between the years 2008 to 2018.

B. Features used and Information Gain

In order to profile the student into various skill-sets, as indicated in Table I, each skill-set has to be defined. Table I illustrates a decomposition of each skill-set based on individual subjects and benchmark assessments.

Table I: The decomposition of skills set based on individual subjects and benchmark assessments along with the weight of each component.

Skill-set	Subject	Weight
Language Skill	English First Language	0.5
	English First Add Language	0.3
	If English is Home Lang	0.2
	NBTAL	0.3
Mathematical Skill	Core Mathematics	0.5
	Mathematics Literacy	0.3
	Additional Mathematics	0.3
	NBTMA	0.2
Computer Skill	Grade 12 Computer Subject	0.7
	NBTQL	0.3
Life Orientation	Life Orientation	1

The first column indicates the skill-set; the second column indicates the Grade 12 subjects and benchmark test associated with the corresponding skill-set in column one; and the third column indicates the weight associated with each subject or benchmark test. The skill-set for learners are derived by allocated points based on the learner's achieved grades for various subjects and benchmarks. The skill is then normalised with respect to the associated weight for each subject.

The language skill-set is primarily composed of English as a Grade 12 subject. Emphasis is put on English since English is the language of instruction at the South African institution we considered. Therefore, a higher language skill-set is allocated to a learner who takes English home language (with subject codes: APENSC; ENA; ENA-HG; ENA-LG; ENANSC; ENA-SG; ENH; ENY) than those who take English as a First Additional Language (ENB-HG; ENB-LG; ENBNSC; ENCNSC;) in Grade 12. [21] provides a list of grade 12 NSC subject codes.

We have also included a National Benchmark Test in Academic Literacy (NBTAL) which attempts to gauge the learner's proficiency to understand academic literacy. The NBTAL

is designed to assess the capacity of first-year students to manage the lingual demands of higher education. This includes language-of-instruction, academic reading and reasoning [22]. Learners who list English as their spoken language at home receive additional points since English is the language-of-instruction at the institution. For example, a student who obtains 70% for English home language (70×0.5), 60% on her NBTAL test (60×0.3), and whose spoken language is English (70×0.2) will receive a language score of $0\geq0.64\leq1$.

Development of early mathematical skills has assumed greater importance in recent times due to its importance on academic achievements and future occupational preparation. Based on previous studies, early mathematical skills are not only associated with later mathematical abilities, but also predictive of other academic aspects such as reading abilities. The Mathematical skill-set primarily comprises of Core Mathematics as a Grade 12 subject. A higher Mathematics skill-set is allocated to a learner who takes Core Mathematics (MATNSC; MAT-HG;MAT) than those who take Mathematics Literacy (MAT-SG; MALNSC) in Grade 12.

We have also included a National Benchmark Test in Mathematics (NBTMA) which attempts to gauge the learner's proficiency in mathematical literacy. The NBTMA is designed to assess the capacity of first-year students to manage basic mathematical demands of higher education. Learners who undertake additional mathematics subjects (such as ADM; ADM-HG; ADMNSC; ADM-SG; ADV; ADV-SG; APMNSC; AVM-HG; MP3NSC; MTA; MTH) at Grade 12 level receive additional points.

The computer skill-set incorporate any computer based subjects (CATNSC; CSC-HG; CSC-SG; CST; CST-HG; CST-SG; CTY-SG; INT; INTNSC) undertaken at Grade 12 level combined with the National Benchmark Test in Quantitative Literacy (NBTQL). The NBTQL assess the learners capacity to problem solve in the higher education context in verbal, graphical, or tabular environments using basic quantitative information.

Life Orientation (LO), is a subject undertaken at schools in South Africa, and focuses on the student's ability to develop his or her full potential in a holistic manner. It is also intended that the learners will develop abilities of making 'good' decisions regarding his or her own health and the environment. LO is also specifically intended to help learners face and cope with problems, such as drug abuse, AIDS, peer pressure, and STDs as well as societal issues and problems such as career choices, work ethic, productivity, crime, and corruption.

The assessment standards in the National Curriculum Statement (2002; 2003) assert that learners are expected to be able to solve or at least manage these problems in constructive ways. Unfortunately, many learners seem to view LO as unnecessary, boring and irrelevant. Furthermore, [23] provides some evidence that LO does not succeed in accomplishing its aims, as laid out in the National Curriculum Statement. Regardless, since all students are required to complete this subject for summative assessment, it is important to consider how LO impacts student success or attrition, and therefore is included as a variable in this study.

Using a distribution of skill-sets to profile the four Science streamlines allows us to visually explore the extent to which each skill-set can be used to represent success in each Science streamline. Figure 1 shows a set of box-and-whisker diagrams indicating the distribution of skill-sets with respect to the associated Science streamline for learners who have successfully obtained an undergraduate degree. In other words, the skill-sets in Figure 1 present the suitable skill-sets required for each streamline.

The description of each box-and-whisker diagram is as follows: the upper and lower outliers have been omitted (more than 3 times of upper quartile or lower quartile); the top whisker represents the greatest value (excluding outliers); the upper quartile indicates a value where 25% of the data is greater; the median indicates a value where 50% of data is greater (the middle of the data set); the lower quartile indicates a value where 25% of the data less than this value; and the bottom whisker represents the least value (excluding outliers).

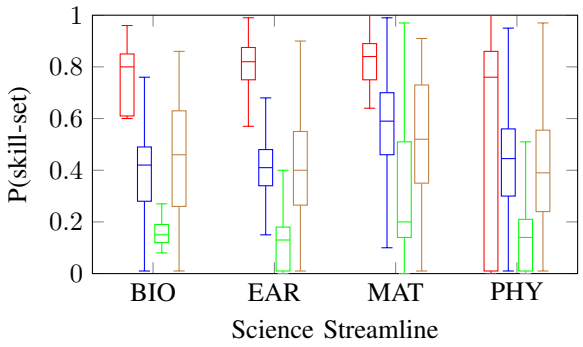

Figure 1: A box-and-whisker diagram indicating the distribution of minimum skill-sets of successful candidates over four Faculty of Science streamlines. KEY: Life Orientation (Red); Mathematical Skill (Blue); Computer Skill (Green); Language Skill (Brown).

Between each of the box-and-whisker diagrams (Figure 1) the distribution of the skill over Science streamlines overlap, although, in some cases there is a clear indication of the threshold level which codes for success in each Science streamline. Four noteworthy observations from Figure 1 include (a) successful students in the Biological, Earth, and Mathematical Sciences have the highest Life Orientation Skill on average; (b) successful students in the Mathematical Science have the highest Mathematical Skill on average compared to the other Science streamlines, followed by Physical Science, Biological Sciences, and Earth Sciences; (c) successful students in the Mathematical Sciences have the highest Computer Skill on average; and (d) successful students in the Mathematical Science have the highest Language skill on average compared to the other Science streamlines.

Although the four skill-sets do provide a guideline in exploring the suitability of learners' to succeed in a particular Science streamline, not all of these skill-sets meaningfully contribute to predicting the learner into a streamline. We can

use feature selection to deduce the most contributing features to predict the class variable. The problem of feature selection is broken up into two components: (a) declaring a mechanism to perform feature evaluation with respect to the class variable (Science streamline), and (b) using this feature evaluator to navigate combinations of features to derive the information loss of using variables subsets of the feature list.

We will use the Information Gain Ranking (IGR) algorithm to perform feature analysis. The IGR algorithm calculates the entropy (information gain) for each feature with respect to the class variable. The entropy, $0 \geq e \leq 1$, where 0 indicates no information gain and 1 indicates maximum information gain. In the next section we will describe the machine learning classifiers used in this paper.

C. Classification and Evaluation

We use the following six off-the-shelf classification models to predict the risk status of a learner: decision trees, k-star, naïve Bayes, support vector machines (SVMs), feed-forward neural networks, and linear logistic regression models.

a) Decision trees: The decision tree algorithm we selected for this task was the $C4.5$ classification model. The $C4.5$ algorithm uses information entropy to build a decision tree based on the ID3 algorithm [24]. The $C4.5$ algorithm recursively selects a feature with the greatest information gain to split the training sample. This intuitively allows the most important feature, with respect to the class variable, to make the 'decisions' from the root down the tree. The $C4.5$ classification procedure implemented in this paper follows the original algorithm by [25].

b) K:* The K* instance-based classifier uses an entropy-based distance function to classify test instances using the training instance most similar to them. The K* implementation used in this paper closely followed the implementation by [26]. Using an entropy-based distance function allows consistency in the classification of real-valued and symbolic features found in our experiments.

c) The Naïve Bayes Model: Perhaps the simplest example of a Bayesian model is the naïve Bayes model (NBM) which has been traditionally and successfully used by many expert systems [4], [27]. The NBM pre-defines a finite set of mutually exclusive classes and assumes that all of the features are conditionally independent given the class label of each instance [28].

The NBM remains a simple, yet highly effective, compact, and high-dimensional probability distribution that is often used for classification problems [29]. The implementation of the NBM follow that of [30].

d) Support Vector Machines: The Support Vector Machine (SVM) classification model incorporates the training data into a non-probabilistic binary linear classifier which separates the classes of the training data by a multi-dimensional hyperplane. Test instances are then mapped on the same space and predicted based on which side of the hyperplane they fall on. Using a kernal trick and the one-verses-all class partitioning, SVMs can be scaled for nonlinear and high-dimensional classification. The SVM implementation used in this paper follows the implementation by [31].

e) Multilayer Perceptron: The multilayer Perceptron used in this paper is a feed-forward neural network which uses Sigmoid functions to represent the nodes and back-propagation to classify instances. The implementation used in this paper follows [32].

f) Linear Logistic Regression Models: The Linear Logistic Regression classification model uses additive logistic regression as mentioned in [33] with added simple regression functions as base learners [27]. The implementation used in this paper follows [34], [35].

All six of these classification models will be evaluated using a confusion matrix [36] and the associated classification accuracy will be provided alongside each model. A 10-fold cross validation scheme will be used [37].

D. Ethics Clearance

The study ethics application has been approved by the Human Research Ethics Committee. The ethics application addresses key ethical issues of protecting the identity of the students involved in the study and ensuring the security of data. The clearance certificate protocol number is $H19/03/02$.

IV. Results and Discussion

In this section we present the results of performing machine learning classification of a learner with varying degrees of Language skill, Mathematical skill, Computer skill, and Life Orientation skill into one of four of the following Science streamline: Biological, Earth, Mathematical, or Physical Science. This section is structured as follows: Section IV-A presents the results of the feature analysis; and Section IV-B presents the classification results.

A. Feature Information Gain

There were 5 features used in this paper. The four skill-sets and the year in which the learner registered for the degree. Using IGR we deduced the contribution of each feature to classify the features as a value from the class variable.

Table II illustrates a ranking of the contribution of each feature. The first column indicates the rank of the feature from most contributing feature (rank 1) to least contributing feature (rank 5); column 2 indicates the entropy value associates with each feature where $0 \leq e \geq 1$ (0 no information to 1 maximum information); and the third column indicates the feature name/description.

Table II indicates that in order of entropy (from highest to lowest), the most contributing features are: (1) the year the learner registered for their degree; (2) mathematical skill; (3) computer skill; (4) language skill; and finally, (5) life orientation skill.

The results indicate that the time the learner registered contributes significantly to predicting the class variable. Table II also indicates that the least contributing features are the language skill and life orientation skill.

Table II: A ranking of the information gain (entropy) for a set of features to predict the learners risk status (class variable). The top three features are highlighted indicating entropy greater than 0.1.

Rank	Entropy (e)	Feature Name
1	0.5181	Year Started
2	0.2507	Mathematical Skill
3	0.1299	Computer Skill
4	0.0452	Language Skill
5	0.0431	Life Orientation

Although the ranking of the feature set using entropy provides a useful framework for feature elimination, the entropy value also indicates the contribution of each feature relative the other features. We see that the entropy values in Table II monotonically decreases logarithmically. The three most contributing features in Table II are highlighted.

B. Classification

In this section we will present the results of the classification algorithms. Figure 2 indicates the result of each of these classifiers to predict the class variable. Classification for instances in the Physical Sciences were the least accurate compared to the other streamlines.

Figure 2a and Figure 2c illustrates the confusion matrix for the $C4.5$ classification model which achieves 58% and 58% accuracy respectively using 10-fold cross validation.

Figure 2b illustrates the confusion matrix for the K* classification model which achieves 57% accuracy using 10-fold cross validation. From all six classification models employed in this paper K* took the least time to build and obtained the worst classification accuracy achieved in this paper. Notable instances that were misclassified include: 30% of Physical Science instances were incorrectly classified as Earth Science.

Figure 2d illustrates the confusion matrix for the SVM classification model which achieves 59% accuracy using 10-fold cross validation. Notable Instances that were misclassified include: 33% of Mathematical Science instances were incorrectly classified as Biological Sciences; 30% of Physical Science instances were incorrectly classified as Earth Sciences. The SVM obtained a 83% accuracy on the Biological Science streamline instances.

Figure 2e illustrates the confusion matrix for the Multi-layer Perceptron classification model which achieves 62% accuracy using 10-fold cross validation the highest classification accuracy achieved in this paper. Furthermore, from all six classification models employed in this paper the Multi-layer Perceptron classification model took the longest time to build. There are no notable misclassification higher than 25% and the multilayer Perceptron seemed to classify the Mathematical Science streamline with 74% accuracy.

Figure 2f illustrates the confusion matrix for the linear logistic regression classification model which achieves 59% accuracy using 10-fold cross validation. With the exception of the Multi-layer Perceptron classification model, the Linear Logistic Regression Model took the longest time to build. There are no notable misclassification higher than 25% and the linear logistic regression classification model seemed to classify the Biological Science streamline with 76% accuracy.

V. Discussion and Conclusion

Although the combined feature set achieves 62% accuracy over the four Science streamlines using the Multi-layer Perceptron, not all the listed features in Table II provide an equal contribution towards correctly classifying the class variable. Furthermore, the achieved accuracy and performance of the SVM reveals the non-trivial relationship between the response and explanatory variables, and the non-linear separability as a property of this data respectively. Despite this, these accuracies are significantly better than a random assignment over the four class values (i.e. random would be 25%).

While the entropy values in Table II monotonically decreases, the entropy of the language skill and life orientation skill are similar, which means that we lose an increasingly smaller entropy with every employed skill-set. This suggests that the language and life orientation skills may not be providing much information to predict the class variable as compared to the mathematical or computer Skill. This is not to imply that the language skill is insignificant, but rather that the entropy gain is mostly dominant in the presence of the computer and mathematical skill-sets.

We note that the Multi-layer Perceptron classification model outperformed the other five models, but provided the longest build-time. The similarity of the skill distribution in Figure 1 between Physical Science and Earth Science must have be the reason for many misclassifications incurred between these streamlines as indicated in Figure 2. This may be the reason why globally the classification for instances in the Physical Sciences were the least accurate compared to the other streamlines in Figure 2.

An objective of this paper is to provide an automated system to predict the suitability of a learner's skill-set to a science streamline in order to calculate the necessary skill improvement required by the learners to be successful in that streamline.

As an example, suppose we would like to calculate the posterior distribution over mathematical sciences using a learner with the following observations: English as spoken home language; with a 57%, 53%, and 53% score in the National Benchmark tests on Academic Literacy, Mathematical Literacy, and Quantitative Literacy respectively; achieving a 60% in Life Orientation; 54% in Core Mathematics; 65% in English Home Language; and 61% in Computer Studies in Grade 12, then the following skill sets can be calculated in Figure 3.

Then, by using the proposed program (available at http://matlabapps.ms.wits.ac.za:9980/webapps/home.), we can deduce that the learner is hypothetically 20% likely to succeed in Biological Sciences; 15% likely to succeed in Earth Sciences; 10% likely to succeed in Mathematical Sciences; and 58% likely to succeed in Physical Sciences.

Even though the learner may be accepted into the mathematical sciences (as their chosen career path), the predictive model calculates the likelihood of success in the mathematical

Actual \ Predicted	BIO	EAR	MAT	PHY
BIO	286	26	86	29
EAR	76	228	30	93
MAT	74	15	314	24
PHY	62	124	83	158

(a) A confusion Matrix describing the performance of the **C4.5** classification model on a set of test data. The **C4.5** classification model achieves **58%** accuracy. 896 correctly classified instances and 722 incorrectly classified ones.

Actual \ Predicted	BIO	EAR	MAT	PHY
BIO	306	39	61	21
EAR	77	240	30	80
MAT	96	11	279	41
PHY	75	127	73	152

(b) A confusion Matrix describing the performance of the lazy **K*** classification model on a set of test data. The lazy K-Star classification model achieves **57%** accuracy. 977 correctly classified instances and 731 incorrectly classified ones.

Actual \ Predicted	BIO	EAR	MAT	PHY
BIO	350	16	49	12
EAR	82	252	18	75
MAT	135	27	238	27
PHY	91	126	52	158

(c) A confusion Matrix describing the performance of the **naïve Bayes** classification model on a set of test data. The lazy naïve Bayes classification model achieves **58%** accuracy. 998 correctly classified instances and 710 incorrectly classified ones.

Actual \ Predicted	BIO	EAR	MAT	PHY
BIO	356	11	60	0
EAR	83	254	49	41
MAT	145	4	278	0
PHY	90	140	75	122

(d) A confusion Matrix describing the performance of the **SVM** classification model on a set of test data. The SVM classification model achieves **59%** accuracy. 1010 correctly classified instances and 698 incorrectly classified ones.

Actual \ Predicted	BIO	EAR	MAT	PHY
BIO	339	15	73	0
EAR	77	236	28	86
MAT	94	9	316	8
PHY	72	107	82	166

(e) A confusion Matrix describing the performance of the **Multilayer Perceptron** classification model on a set of test data. The Multilayer Perceptron classification model achieves **62%** accuracy. 1057 correctly classified instances and 651 incorrectly classified ones.

Actual \ Predicted	BIO	EAR	MAT	PHY
BIO	324	22	74	7
EAR	71	233	25	98
MAT	91	26	282	28
PHY	68	112	74	173

(f) A confusion Matrix describing the performance of the **linear logistic regression** classification model on a set of test data. The linear logistic regression classification model achieves **59%** accuracy. 1012 correctly classified instances and 696 incorrectly classified ones.

Figure 2: A set of confusion matrices describing the performance of several classification models on a set of test data. Each accuracy of the classification models are indicated along with the correctly and incorrectly classified instances.

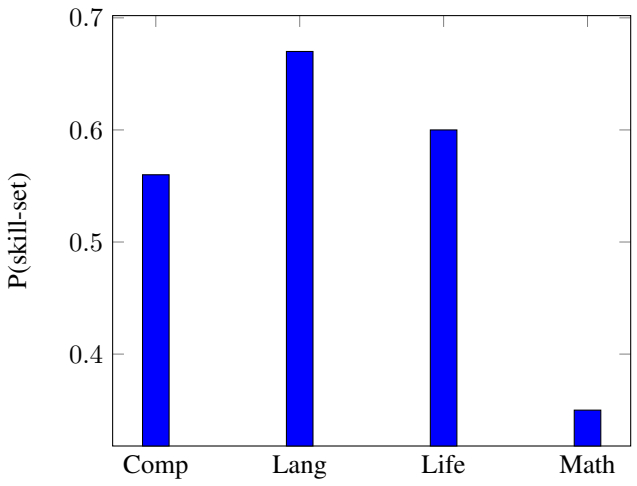

Figure 3: A bar graph showing the skill distribution for a learner.

science as dismal. In light of the skill profile required to succeed in the mathematical sciences (given in Figure 1), the learner needs to consider improving their current mathematical and life orientation skills to be successful. Herein lies the power of this tool in student support since this model can be used to identify skill gaps that the learner can fill to increase their chances of success. Faculty interventions and student support initiatives can thus be used effectively to support the development of the learners skill deficit.

This paper opens a discussion of the power of using skill-sets to identify knowledge gaps which can present an obstacle to obstacle to students success. Future avenues of research can (a) explore the impact of the highly ranked features in Table II; (b) model a more sophisticated representation of each skill-set using more enrolment observations (possibly even qualitative) from the learner, and perhaps (c) model the acquisition of skills based on the learner performance and intellectual development over time so we can acquire a firm relationship between the student and the Faculty streamline.

ACKNOWLEDGEMENTS

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant numbers: 121835 and 121960). The first author gratefully acknowledges the support of Dr. Nishana Parsard for her valuable and constructive suggestions during the development of this research project.

REFERENCES

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

[2] T. Nkambule, "Against all odds: the role of'community cultural wealth'in overcoming challenges as a black african woman: part 2: being and belonging in south african higher education: the voices of black women academics," *South African Journal of Higher Education*, vol. 28, no. 6, pp. 1999–2012, 2014.

[3] E. L. Usher, C. J. Ford, C. R. Li, and B. L. Weidner, "Sources of math and science self-efficacy in rural appalachia: A convergent mixed methods study," *Contemporary Educational Psychology*, vol. 57, pp. 32–53, 2019.

[4] 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.

[5] J. Nicholas, L. Poladian, J. Mack, and R. Wilson, "Mathematics preparation for university: entry, pathways and impact on performance in first year science and mathematics subjects," *International Journal of Innovation in Science and Mathematics Education (formerly CAL-laborate International)*, vol. 23, no. 1, 2015.

[6] M. N. Giannakos, I. O. Pappas, L. Jaccheri, and D. G. Sampson, "Understanding student retention in computer science education: The role of environment, gains, barriers and usefulness," *Education and Information Technologies*, vol. 22, no. 5, pp. 2365–2382, 2017.

[7] S. Dukhan, A. Cameron, and E. A. Brenner, "The influence of differences in social and cultural capital on students' expectations of achievement, on their performance, and on their learning practices in the first year at university." *International Journal of Learning*, vol. 18, no. 7, 2012.

[8] A. Msimanga, P. Denley, and N. Gumede, "The pedagogical role of language in science teaching and learning in south africa: A review of research 1990–2015," *African Journal of Research in Mathematics, Science and Technology Education*, vol. 21, no. 3, pp. 245–255, 2017.

[9] L. Stoffelsma and W. Spooren, "The relationship between english reading proficiency and academic achievement of first-year science and mathematics students in a multilingual context," *International Journal of Science and Mathematics Education*, pp. 1–18, 2018.

[10] E. Prinsloo, "Implementation of life orientation programmes in the new curriculum in south african schools: perceptions of principals and life orientation teachers," *South African Journal of Education*, vol. 27, no. 1, pp. 155–170, 2007.

[11] C. Beaulac and J. S. Rosenthal, "Predicting university students' academic success and major using random forests," *Research in Higher Education*, pp. 1–17, 2019.

[12] R. Ajoodha, A. Jadhav, and S. Dukhan, "Forecasting learner attrition for student success at a south african university," in *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.

[13] A. P. Rovai, "In search of higher persistence rates in distance education online programs," *The Internet and Higher Education*, vol. 6, no. 1, pp. 1–16, 2003.

[14] J.-H. Park, "Factors related to learner dropout in online learning." *Online Submission*, 2007.

[15] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert systems with applications*, vol. 33, no. 1, pp. 135–146, 2007.

[16] F. Araque, C. Roldán, and A. Salguero, "Factors influencing university drop out rates," *Computers & Education*, vol. 53, no. 3, pp. 563–574, 2009.

[17] M. Tan and P. Shao, "Prediction of student dropout in e-learning program through the use of machine learning method," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 10, no. 1, pp. 11–17, 2015.

[18] Z. J. Kovacic, "Early prediction of student success: Mining students enrolment data," in *In SITE 2010: Informing Science+ IT Education Conference*, vol. 10, 2010, pp. 647–665.

[19] R. Woodman, "Investigation of factors that influence student retention and success rate on open university courses in the east anglia region," *Unpublished M. Sc. Dissertation), Sheffield Hallam University, UK*, 2001.

[20] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting students'performance in distance learning using machine learning techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004.

[21] "Nsc subject codes for grade 12 learners," https://wcedonline.westerncape.gov.za/documents/SeniorCertificate/SubjectsToStudy.html, accessed: 2020-04-23.

[22] A. Cliff, "The national benchmark test in academic literacy: how might it be used to support teaching in higher education?" *Language Matters*, vol. 46, no. 1, pp. 3–21, 2015.

[23] A. Jacobs, "Life orientation as experienced by learners: a qualitative study in north-west province," *South African Journal of Education*, vol. 31, no. 2, 2011.

[24] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree id3 and c4. 5," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2, pp. 0–0, 2014.

[25] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

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

[27] R. Ajoodha, R. Klein, and B. Rosman, "Single-labelled music genre classification using content-based features," in *2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*. IEEE, 2015, pp. 66–71.

[28] R. Ajoodha and B. Rosman, "Learning the influence structure between partially observed stochastic processes using iot sensor data," in *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[29] R. Ajoodha and B. Rosman, "Tracking influence between naïve bayes models using score-based structure learning," in *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*. IEEE, 2017, pp. 122–127.

[30] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, 1995, pp. 338–345.

[31] Y. EL-Manzalawy, "Wlsvm," 2005, you don't need to include the WLSVM package in the CLASSPATH. [Online]. Available: http://www.cs.iastate.edu/~yasser/wlsvm/

[32] R. Arora, "Comparative analysis of classification algorithms on different datasets using weka," *International Journal of Computer Applications*, vol. 54, no. 13, 2012.

[33] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," Stanford University, Tech. Rep., 1998.

[34] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine Learning*, vol. 95, no. 1-2, pp. 161–205, 2005.

[35] M. Sumner, E. Frank, and M. Hall, "Speeding up logistic model tree induction," in *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Springer, 2005, pp. 675–683.

[36] K. M. Ting, "Confusion matrix," *Encyclopedia of Machine Learning and Data Mining*, pp. 260–260, 2017.

[37] P. Zhang, "Model selection via multifold cross validation," *The Annals of Statistics*, pp. 299–313, 1993.