Evaluation of Student Skill-Sets as predictors of Success at Higher-education Institutions

Luyanda Makhoba[1], Ashwini Jadhav[2], Khanyisile Sixhaxa[3,*], and Ritesh Ajoodha[4]

[1] School of Computer Science and Applied Mathematics,
University of the Witwatersrand, Johannesburg, South Africa
luyanda.makhoba1@students.wits.ac.za
[2] Faculty of Science
University of the Witwatersrand, Johannesburg, South Africa
ashwini.jadhav.wits.ac.za
[3] School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg, South Africa
khanyisile.sixhaxa1@students.wits.ac.za
[4] School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

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. The best performing prediction model proved to be the Random Forest classifier which gave an accuracy of 95 %. 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 a student's ability. With the advent of machine learning, this study also 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.

Keywords: Data mining, Student attrition, Classification models, Skill-set, Machine Learning

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 rates 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 [11,2]. 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 a discretized approach where a subject mark is given a point score if it is within a certain range and we see the loss of information and causes student ability misrepresentation. As an alternative to the flawed conventional APS entrance criteria, we propose a more bespoke model to represent a student's 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 [13] and we we made adjustments to the original model to form the conceptual framework that this research paper encompasses, illustrated in Figure 2. Based on this underlying scheme, we define the composition of each of the skill-sets, and 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 subsequently use this to predict their success in a particular Science degree streamline. The four Science streamlines are partitioned as: Biological, Earth, Mathematical and Physical Sciences. The skill sets are composed of high school results, National Benchmark Tests(NBT) and biographical data. 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. This can model be used by university administration staff, bursaries and scholarship sponsors to identify academically challenged students and more specifically, identify the additional assistance the student will require based on the skill set characteristic that fall short of. Different forms of support can then be provided to the student to improve their prospects of success [11]. 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 [8]. A range of machine classification models were chosen and trained on the data, namely Decision Trees, Decision Tree Naive Bayes, Multi-layer Perceptron, K-star and SVM. In the next section, we will provide a in-depth review of recent literature on the evaluation of students' skill-sets and how they are used as predictors in higher education institutions.

Table 1: Description of abbreviations used in this research paper

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

2 Related Work

A student's entry academic characteristics influences their success, failure or withdrawal from a particular degree [5]. Based on a student's 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 programmes 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 the students are guided into fields they are best suited for [12]. 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 [14]. 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 school student attrition rate is due to the misalignment of student academic skills and degree streamline? Can prospective 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? In this paper we present 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.

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 [5]. 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 [1].

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.[7,**?**] 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 [6].

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 [15].

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 [9]. Table 2 compares the approaches and performances of the related work investigated in line with this study.

Table 2: Literature Resources - Key contributions

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%

2.1 Conceptual Framework

Figure 1 depicts conceptual framework that our structuring is based on. This an extension of a model proposed by Tinto[13]. 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's attributes, the type and quality of pre-college schooling along with family background influences the pre-college marks of the student. In this research, we link the student's marks with a related 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.

2.2 Skill-Sets Rationale

The readiness theory states that a prospective students readiness can only be evaluated by understanding the demands and expectations of the environment and field they are heading into [6]. The use of skill-sets is in alignment with this notion as the composition of skill-sets is based on key qualities necessary for

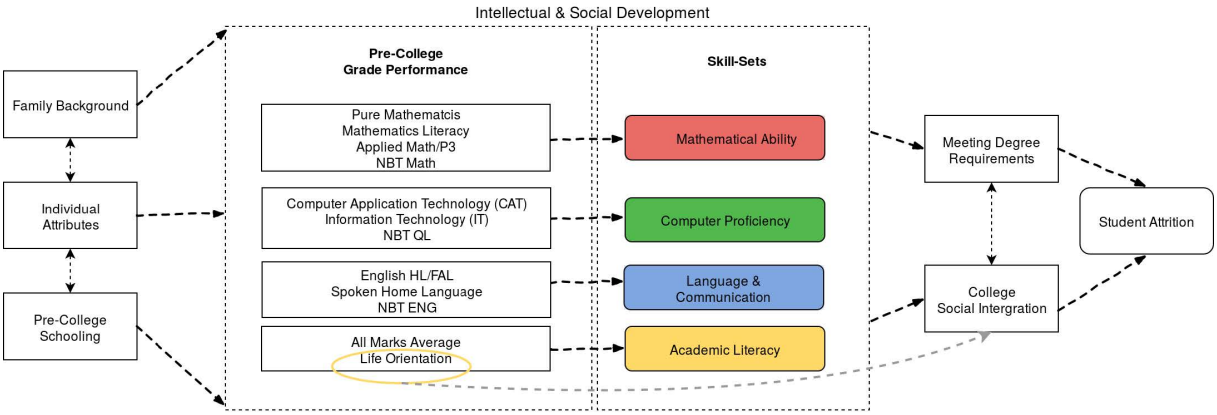

Fig. 1: Applied Conceptual Framework

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 [6]. 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. Mathematical competency 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, the number of languages studied, looking at the marks of the home language studied by the student, weighing the mark based on which level of English was studied in high school and the NBT English mark. The ability of a 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 [4]. As part of this research we use language as one of the skills evaluated because 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 is ones ability to write accordingly and effectively express themselves. Articulation, the ability to effectively explain information and Academic Literacy skills are essential for success in university , because of the high level of academic assessments student undergo. [3]. We incorporate life orientation in this skill-set as this has social aspects taught to learners that influences their academic ability at university

2.3 Methodology

The intended purpose of this research paper is to evaluate the use of students 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.

From the exhaustive list of different machine learning paradigms, in this research we make use of: Bayesian, decision trees, deep learning, instance based, ensemble 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 2.1 describes how the data used was collected along with a comprehensive description of the initial structuring of data . Furthermore, we describe steps undertaken to perform the pre-processing to our data structure to remove unwanted data instances.

The subsequent subsection discusses the composition of student skill-sets from the learner's pre-college attributes, this is then followed by an analysis of the contribution each of the skill-sets provide when used as a predictor for students success. Section 2.2 outlines the generic implementation of the machine learning classification models and an overview of the evaluation metrics used.

Data Analysis and Pre-processing In this paper, the data of students enrolled 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). The data, consisted of the students pre-university entrance results, demographics and with final outcomes/ results (graduation or failure/attrition).

The data received was presented as two types of table sets. The first set had students 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 are 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[10]. We ended up with 342 instances for each class.

Table 3: Breakdown of extracted student table csv

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

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 Mathematics 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.

The language communication ability places emphasis on English, as this is the medium of communication at South African universities 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.

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 Computational Applied Technology therefore it has been given a slightly higher weighting.

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.

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

Table 4: Skill-set mark composition weighting

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

Applying Classification Models In this study we utilized 6 off the shelf machine learning classifiers, namely Bayesian Network, Decision Trees, Multilayer Perceptron, Bagging, Support Vector Machine and K-Star, using 10 fold cross-validation with each classifiers.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. These classification models will be evaluated using a confusion matrix and consequently the accuracy of the model.

2.4 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.

3 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 3.1 the performances of these models are presented using confusion matrices. Subsection 3.2 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 3.3 Provides an analysis of the information gain provided by each of the skill-sets when making classifications.

3.1 Classification Models Predictions

Figure 3 illustrates the confusion matrices of the six selected classification models Figure 3(a) illustrates the confusion matrix of the Random Forest classifier, which achieved accuracy of 95% which makes it best performing classifier. Figure 3(b) illustrates the confusion matrix of the K-Star classifier, which achieved accuracy of 93% being the second best performing classifier. Figure 3(c) illustrates the confusion matrix of the Naive Bayes classifier, which achieved accuracy of 78%, which makes it the worst performing classifier. Figure 3(d) illustrates the confusion matrix of the Multi-Layer Perceptron classifier, which achieved accuracy of 91%, which makes it the best performing classifier with the exception of Random Forest, K-Star and Logistic Regression. Figure 3(e) illustrates the confusion matrix of the J-48 classifier, which achieved accuracy of 88% being the worst performing classifier with the exception of Naive Bayes. Figure 3(f) illustrates the confusion matrix of the Logistic Regression classifier, which achieved accuracy of 92%, which is the best performing classifier with the exception of Random Forest and K-Star.

The Random Forest classifier was the best performing classifier due to the fact that it is ensemble model that consist of decision trees, so scaling of data instances doesn't matter, to avoid overfitting they sample random subspaces in the data coupled with bagging. The Naivee Bayes classifier was the worst performing classifier because of the inaccurate data instance attribute association probability estimate.

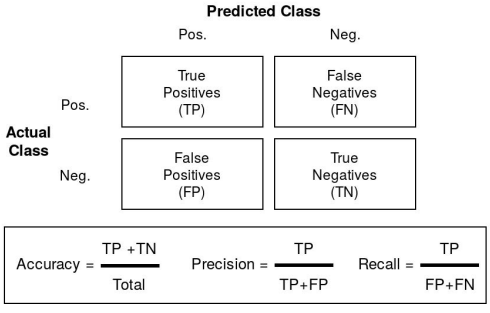

Fig. 2: Confusion Matrix breakdown

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

(a) Random Forest - 95% Accuracy, Correctly identified 731/762 labels

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

(b) K-Star - 93% Accuracy, Correctly identified 719/762 labels

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

(c) Naive Bayes - 78% Accuracy, Correctly identified 592/762 labels

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

(d) Multi-Layer Perceptron - 91% Accuracy, Correctly identified 699/762 labels

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

(e) J-48 Algorithm - 88% Accuracy, Correctly identified 674/762 labels

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

(f) Logistic regression - 92% Accuracy, Correctly identified 707/762 labels

Fig. 3: A set of confusion matrices describing the performance of several classification models on a set of test data. Each classification model's accuracy is indicated along with the correctly and incorrectly classified instances.

3.2 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.

3.3 Skill-set attributes across streamlines

In this subsection we investigate how the skill-set values are distributed and differ across the different degree streamlines. Figure 4 depicts the box and whisker plots

Table 5: Entropy ranking 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

of students a) who have qualified and b) students who have failed to qualify and obtain their degree.

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.

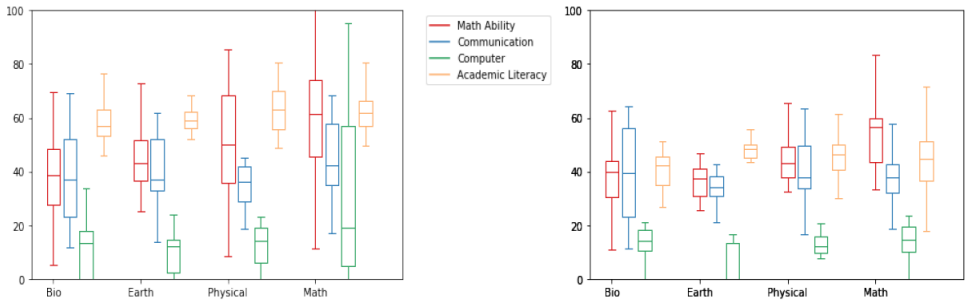

Fig. 4: Box and Whisker Plot of Skill-set distribution by Degree Streamline

4 Conclusion

In this work we used learner skill-sets to represent their academic ability when entering university. With this model approach investigated in this research paper we attest that it is a viable alternative to the conventional APS entrance criteria. 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

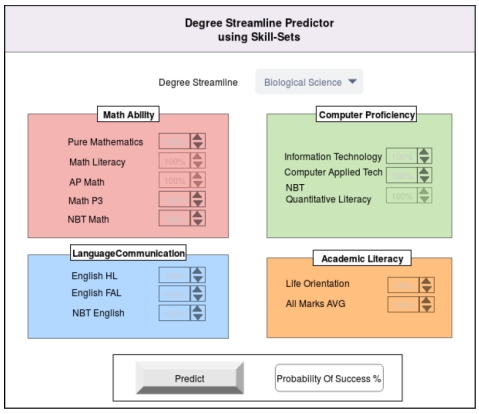

Fig. 5: App Prototype

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, Random Forest, as the back end functioning of a proposed application prototype shown in figure 8. This application is aimed at prospective university students intending to pursue a career in the sciences. Limitations faced in this paper were that there was a period in time where there was a political unrest in the country that affected students academic performance and also the discretization of classes can lead to misclassification because realistically classes can be modelled continuously.

Acknowledgement

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

References

1. Agarwal, S., Pandey, G., Tiwari, M.: Data mining in education: data classification and decision tree approach. International Journal of e-Education, e-Business, e-Management and e-Learning **2**(2), 140 (2012)
2. Ajoodha, R., Dukhan, S., Jadhav, A.: Data-driven student support for academic success by developing student skill profiles. In: 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC). pp. 1–8. IEEE (2020)
3. Andrews, D., Osman, R.: Redress for academic success: possible'lessons' for university support programmes from a high school literacy and learning intervention. South African Journal of Higher Education **29**(1), 354–372 (2015)
4. Barac, R., Bialystok, E.: Bilingual effects on cognitive and linguistic development: Role of language, cultural background, and education. Child development **83**(2), 413–422 (2012)
5. Campbell, P.F., McCabe, G.P.: Predicting the success of freshmen in a computer science major. Communications of the ACM **27**(11), 1108–1113 (1984)
6. Conley, D.T.: Redefining college readiness. Educational Policy Improvement Center (NJ1) (2007)
7. Cortez, P., Silva, A.M.G.: Using data mining to predict secondary school student performance (2008)
8. Maddikunta, P.K.R., Pham, Q.V., Prabadevi, B., Deepa, N., Dev, K., Gadekallu, T.R., Ruby, R., Liyanage, M.: Industry 5.0: A survey on enabling technologies and potential applications. Journal of Industrial Information Integration p. 100257 (2021)
9. Pandey, M., Taruna, S.: Towards the integration of multiple classifier pertaining to the student's performance prediction. Perspectives in Science **8**, 364–366 (2016)
10. Reddy, T.G., Bhattacharya, S., Maddikunta, P.K.R., Hakak, S., Khan, W.Z., Bashir, A.K., Jolfaei, A., Tariq, U.: Antlion re-sampling based deep neural network model for classification of imbalanced multimodal stroke dataset (2020)
11. Scott, I., Yeld, N., Hendry, J.: A case for improving teaching and learning in south african higher education. Higher education monitor **6**(6), 1–83 (2007)
12. Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., Schmidt-Thieme, L.: Recommender system for predicting student performance. Procedia Computer Science **1**(2), 2811–2819 (2010)
13. Tinto, V.: Dropout from higher education: A theoretical synthesis of recent research. Review of educational research **45**(1), 89–125 (1975)
14. Wilson-Strydom, M.: 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)
15. Zohair, L.M.A.: Prediction of student's performance by modelling small dataset size. International Journal of Educational Technology in Higher Education **16**(1), 1–18 (2019)