UNIVERSITY OF THE WITWATERSRAND

WITS
UNIVERSITY

School of Computer Science and Applied Mathematics
Faculty of Science
HONOURS RESEARCH REPORT

A Programme Recommendation Engine to Improve Student Placement at a South African Higher-Education Institution

Tasneem Abed

Supervisor(s): Dr. Ritesh Ajoodha and Dr. Ashwini Jadhav

October 2019, Johannesburg

DECLARATION
UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
SCHOOL OF COMPUTER SCIENCE AND APPLIED MATHEMATICS
SENATE PLAGIARISM POLICY

I, Tasneem Abed, (Student number: 1408535) am a student registered for BSc with Honours in Big Data Analytics in the year 2019.

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: tassabed

Signed on 18th day of October , 2019 in Johannesburg.

A Programme Recommendation Engine to Improve Student Placement at a South African Higher-Education Institution

Tasneem Abed

Abstract

There is a growing concern over the low pass rates of students in the Science Faculty in a higher education institute in South Africa. The Admission Point Score (APS) used to place students into programmes may appear to have good discretion in gauging students aptitude, but the reality is that between 2008 and 2015, about 50% of students who met the APS requirements for a Science programme failed to meet the requirements to pass. This report attempts to build a recommendation engine that will advise students on their academic trajectory for a chosen programme based on features suggested by the Tinto (1975) framework. It will also rank these features according to their contribution to the models predictability. The results show that classification models from various archetypes of machine learning have high accuracy in predicting whether a new student is likely to qualify in 3 years, more than 3 years or fail to meet the requirements. This proves to be a better mechanism for placing students than the APS. Individual attributes, which do not include any high school marks, have the highest contribution to the models' accuracy. This research argues that a more complex view of student placement will improve the faculties success rates. Additionally, an engine that will allow students to get an idea of how their academic career will play out will optimise their outcome and make for a higher throughput per year.

I. Introduction

Throughout the world, higher education has become accepted as a vital key for national development in the context of the knowledge community and globalization. In South Africa, this places a big concern on the output of the higher education sector in terms of the number of graduates as well as quality of graduates. Of most concern is the mismatch between the output of the higher education sector and the economic needs of the country that exist [1]. It is important that the performance of this sector is critically and constantly assessed. Performance can be measured by a number of features. In this research we will focus on performance of students and throughput of graduates.

Narrowing the scope of the higher education sector down to universities, failure rates in certain faculties are quite alarming, particularly in the science faculty. In an investigative report on higher education monitoring, it was reported that only 50% of students who entered into a 3-year mathematical science degree graduated within 5 years, i.e. with an additional 2 years to make up modules previously failed. The statistics for other fields are similar. The highest attrition rate occurs at the end of the first year of study with about 29% of first year students [1].

There are numerous factors that could lead to a student dropping out or failing and identifying these factors is no simple task. The academic performance of students is influenced by many factors, however the influence of their performance in high school is heavily weighted in the admission process. Admission boards use the Admission Point Score (APS) as a means of offering students a place in academic programme. The APS is a weighted calculation based on the symbols received for each Grade 12 subject. This means the number of points earned for receiving 50% in a subject is the same as the amount received for 59%. National Benchmark Tests (NBT) may also be looked at to supplement the board's decision. The APS may provide an indication of students performance, however, between the years 2008 and 2015 nearly 50% of students who obtained an adequate APS required for a programme failed to complete it [2]. This reflects one of the downfalls of this system in that it lacks qualitative discretion i.e. the content of the school subjects is not taken into account. For example, a student may achieve a high APS by obtaining distinctions for all their subjects and thus qualifying to enrol in a degree in the Sciences. However, these Matric subjects may not be Science subjects but Art or Language based subjects instead. In this case, given that there is a positive relationship between the participation in Science related activities in high school and student achievement in Science [3], these students may not have the adequate skills to succeed in a Science degree despite meeting the APS requirements.

At the root of every student's academic trajectory is the decisions they made that led them to register for their courses. These decisions carry a lot of weight for their future and should not be taken lightly. From around Grade 11, students begin to seriously think about what they are going to be doing after school. The average age of a Grade 11 student is 17 years. The decision to attend university and what to study there is a big decision for teenagers at this age and there may be a lack of maturity in the governing of these decisions. This is where good academic advisory can be the difference between a student failing a course and passing it, or a student passing a course but changing to another that

they discover that they are more interested in.

Good academic advisory for every student could potentially have a positive effect on failure rates as students will be making more informed decisions about their academic trajectories. Students should know, prior to the commencement of their degree, whether they are more likely to pass in minimum time, pass in more than minimum time or fail to meet the minimum requirements. Of course, students apply for courses that they have an interest in despite the difficulty level. However, if they are advised that passing may be difficult for them, they do not have to abandon the course but rather know that they may have to work harder to pass than expected. It is also vital that they are aware of all the different available courses prior to registering for a course so that they do not waste a year doing something only to change to another course that they only hear about later. However, how do we quantify 'good' advisory? Is it enough to glance over a student's high school results and develop an idea of their expected success? Are their high school results sufficient indicators of success? Should advisory be more concrete in the form of numbers i.e. probabilities, percentages and statistics?

In this study, a dataset containing the undergraduate course registration and Matric results from 2008 to 2018 will be put through machine learning models in order to predict the success rate of a student given their profile. The conceptual framework proposed by Tinto (1975) [4] is adopted to develop a methodology to predict the success of completing a degree. In particular, six models with different structures and advantages will be implemented and their performances compared. Profiles of students are made up as a combination of features according to the three categories in the Tinto (1975) framework: Background attributes, Individual attributes and Pre-University attributes. These include their Matric marks, NBT marks and features such as quintile of the high school they attended, the setting of their high school, such as urban or rural and their age at the commencement of first year, as seen in Table I (where Mathematics is shortened to Math). An analysis of the results will be conducted to see which of the Tinto (1975) categories contribute the most to the predictability of the models and if there are any unexpected results from the ranking of the information gain of the features. In order to convey the expected success rate of individuals in specific courses meaningfully, a graphical user interface will be built that will display to the student how their features effect their success rate of a chosen programme.

This research will, in general, contribute an indication of how profile features of students influence their rate of success and provide a ranking of importance of features through information gain. Additionally, the need for a more complex mechanism to determine student placement will be argued. These will be data driven contributions and will underpin a user interface that will assist students in making decisions about their academic trajectories.

In section II, we will provide a literature review in order to understand the trends, models, frameworks and results that have been produced in the research field of educational data mining thus far and how it influences this research. Section III will outline the methodology of the pre-processing of the data, the experiment set up of the models and the evaluation thereof. This is followed by the results and analysis in section IV and finally the concluding remarks.

II. Background and Related Work

In order to have a full understanding of educational data mining, and the techniques and processes applied in building such an engine, it is vital to review and understand the context of the research field by delving into the current state of the field. In this section, we will present the background necessary to build a student advisory system for the Science Faculty in a South African context for a research intensive South African University.

Tinto (1975) proposed a widely cited conceptual framework of the student attrition process. In this model, three groups of characteristics, namely Background, Individual and Pre-College/Schooling, are interrelated and expected to influence a student's determination into achieving the goal of graduating [5]. Higher grade performance and intellectual development is achieved through commitment to the goal which leads to academic integration and reduces the likelihood of dropping out.

A. Educational data mining

The final 3 years of high school in South Africa (Grades 10 - 12), requires learners to take 4 mandatory subjects, namely English (Home Language or First Additional Language), Mathematics (core or literacy), a second approved language (first or second additional language) and Life Orientation. Additionally, a minimum of 3 other subjects must be chosen. These include Geography, Life Sciences, Physical Sciences, History and Music amongst others. In the final year of high school, Matric, learners write national examinations that are set by the Department of Basic Education. A combination of the marks of these examinations and the marks earned during the final school year are used to calculate an Admission Point Score (APS), with the highest score for a single subject being 8 [6].

Universities in South Africa use the APS of students as a criterion for admissions. Some programmes look at percentages of specific subjects to supplement their admission protocols. Computer Science looks at mathematical ability as a primary predictor of success and thus weights Matric Mathematics marks more heavily for admission [7]. This is backed up by numerous studies that show a clear link between mathematical aptitude and programming such as one done by Byrne *et al.* [8]. In any Mathematical Science degree, mathematical aptitude would naturally be the primary concern of an admission board as it is directly correlated to the subject matter of the degree's content. If a student has poor results

in Matric Mathematics, the chances of them succeeding in a tertiary setting is minimal. In fact, according to Campbell *et al.* [9], many universities throughout the world use high school Mathematics results to select their students. However, this may not be the only significant factor to consider. Success in a range, or subset, of high school subjects should be considered.

The language of instruction of the university (English) plays an important role in the success of a student based on the student's comfort with the medium. Although it may be overlooked when considering admission into a Mathematical Science programme, the ability of a student to comprehend the content that will be provided to them should also be a primary concern, especially given the South African context where many learners are not native English speakers. English performance in high school is a quantified measure of comfort with the English language which is coupled with their NBT mark. An investigation into the belief that language ability influences success at university [7] found that achievement in high school language courses is a better predictor of success than mathematics. Through information gain ranking, we will investigate the contribution of the Matric English marks and the Quantitative Literacy NBT mark i.e. NBTQL.

Biographical profiling of students at university is practised very early into the student-university relationship. Before a student is offered a place, the university already has a wealth of information about them, including their home language, race, gender and home province. These features are considered in an investigation into at-risk undergraduate profiles [10]. Gender is a factor that has shown to provide some insight into persistence. Females are less likely to persist in scientific majors, however this is not indicative of academic achievement or potential [9]. A distribution of home languages of students in Mathematical Science degrees shows that 38% of students have English as a home language. Although this was the modal home language, it is very low when considering the implications of English ability as discussed previously [10]. The home province of a student and their home language are directly linked. It is also difficult to consider home province as there may be less students who come from locations further from the university, thus creating data limitations.

Race is a sensitive topic, especially in a South African context. Nevertheless, a study on the influence of race on student performance in Mathematical Science degrees showed that 63% of students had an associated Black race description and, alarmingly, 71% of at-risk profiles also had an associated Black race description. These studies have produced significant results which show that it is worth taking into account biographical features when looking at success rates and not just high school marks. However, these features may be easily misinterpreted in the context of this research project as the future is being predicted here instead of an analysis of the past. These features may quickly lose their statistical importance over a few years which would depreciate the quality of the recommendation engine and thus will not be used.

Perhaps an even more worthwhile category of features to use is "abstract" abilities such as comprehension skills, memorization skills, programming skills, mathematics skills and inferential thinking skills which are defined as core features or characteristic skills that a student needs to possess in order to succeed in the course [11]. No insight is given as to why these features are used instead of high school marks, however the study can be used to compare the effect of using different types of features for the same purpose. Unfortunately for this research project, there is no data on certain abstract abilities such as programming skills, computer skills or inferential thinking skills available from the university, but NBT results may give an indication of mathematics and comprehension skills.

Academic advisory is a key step in any student's academic journey. Insufficient advising is not an uncommon practice especially in the world of distance education where students do not have one on one interaction with advisors and academic staff. With numerous courses to select from, a student may not be sure of their interest in a course solely based on its title [11]. All the above mentioned feature engineering and importance provides a solid base for building a system that will help students to understand their academic trajectories and make informed decisions to optimize their studies to be more feasible, worthwhile and rewarding.

B. Mathematics and techniques used in educational data mining

Statistical analysis is a method used for prediction and correlations. In more recent years, machine learning has been developed into a field of its own that encompasses a plethora of algorithms that serve the purposes of clustering and prediction. This section will detail the mathematical approaches used in the discovery of results related to education as well as machine learning techniques used in building predictive and recommendation engines in the context of education.

Statistical analysis is an approach taken towards the collection and interpretation of data. In a study done in 1984 that sought to predict the success of first years in a Computer Science major, a statistical analysis approach was taken to determine if there was a significant difference between students of Computer Science, Engineering, Other Sciences and Others, along with any entrance variables, to determine which combinations of entrance variables could predict the group of a student [9]. Common statistical measures such as the mean, variance and standard deviations are calculated for each group and entrance variables. One sided t-tests were used to compare the Science group versus the Other group to see if the Science group had significantly higher mean values. Statistical analysis was not sufficient to justify that the difference found between the groups were useful or important, so discriminant analysis was done instead.

Discriminant analysis is a statistical classification technique that is used to assign a student a group. Wilks' lambda discriminates between 2 groups on the basis of multivariate t-tests and gave a 68.4% accuracy which proved to be the best classification method.

Mathematical approaches taken to tackle the problem of predicting the most suitable programme for a student, in order to recommend it to them, revolve around both supervised and unsupervised machine learning algorithms. In a study by Aly *et al.* [12] conducted in Egypt, the K-means clustering algorithm is applied to divide student records into clusters based on similarities in marks. After the cluster to which a student belongs has been identified, various decision tree algorithms are applied in order to recommend a department with the highest success rate for the student. These algorithms include the ID3, C4.5 (also called J48) and CART algorithms. Of these, J48 proved to be the most efficient and robust. This is because it represents results of research in machine learning that traces back to the ID3 system and thus has been taken as a point of reference for the development and analysis of novel proposals. It provides good classification accuracy and is very efficient [13].

Support vector machines (SVMs) have been shown to give good generalization performance on a variety of problems, however, can suffer from slow training and high complexity. The Sequential Minimal Optimization (SMO) algorithm is an advancement of SVMs with better scaling properties. It makes use of an analytic quadratic programming step [14]. In a study to predict student performance, a number of machine learning algorithms from different paradigms with different mathematical properties were tested to see which would perform best. Results showed a Multilayer Perceptron (MLP) had the highest accuracy in predicting grades [15]. This approach is able to provide some insights into different machine learning algorithms through comparison.

As shown, the mathematical approaches used in this field vary considerably and no one approach is proven to be the best. Although clustering is not a concern in this research, the success of decision tree algorithms, such as the J48 algorithm, for predicting successful programme makes it an algorithm to consider for finding the most suitable programme to recommend to a student. It is also an algorithm in which information gain is a key calculation. To identify which of the Tinto (1975) categories prove to be more indicative of success, we can look at the information gain and see how they fair against each other. This may prove that high school marks are not the most influential attributes. Other models may provide higher accuracies and therefore numerous models from different machine learning paradigms will be implemented and compared.

III. Research Methodology

This research aims to develop a recommendation engine to help students make better informed decisions about their academic trajectory. This will be done by modelling trends in success rates based on Matric marks and biographical profiles of previous and current students, as per the Tinto (1975) framework. The failure rates in the Science field poses the question of whether acceptance criteria needs to be adjusted. This research will also investigate the important features to look at when accepting students by looking at features with high information gain.

These aims will be achieved through the following objectives:

- Investigate the current best features that indicate success.
- Implement and compare the performance of 6 structurally different machine learning models.
- Identify features with the highest information gain and compare them against the current features looked at for accepting students.
- Analyse and discuss the data generated from experiments.
- Build a graphical user interface for students to be able to obtain their expected success rates in courses they are interested in.

A. Data

Two datasets, one containing the Undergraduate Science Course Registration data and another containing the corresponding Undergraduate Science Matric results, were obtained from the Academic Information and Systems Unit (AISU) at the University and comes with high confidence of accuracy. The datasets contain academic information for students from 2008 to 2018 and each new record has a new subject that the student did in either university or Matric, which makes for a lot of repeated data. The data also contains an indication of the final outcome of the student and the number of years taken to obtain that outcome.

B. Ethical Clearance

The use of the data received ethical clearance from the board of ethics at the University. Participants involved in this study were students from the research intensive University in South Africa. The ethics application for this study has been approved by the University's Human Research Ethics Committee (Non- Medical). The application addresses the security of the identities and data of all the participants. The protocol number for the clearance certificate is *H19/08/01*.

C. Preprocessing

The two datasets were joined to create a single dataset where each record represented a student with all their information within one record. Students who joined the University in either 2017 or 2018 would not have surpassed the minimum time to complete their degrees by the end of this research, so those students were removed from the dataset. The target variable contains three possible values that each student can take: *QualMin*, *Qualified* and *Failed*. *QualMin* represents a student who qualified in three years, *Qualified* represents a student who qualified in more than three years and *Failed* represents a student who was unable to obtain their degree. The international status, home province and school setting

variables were then cleaned. International students received an 'IS' value if they were not South African and those who were received the 'ZAF' value. The school setting represents whether the high school attended by the student was in a rural or urban setting. Not all Matric subjects were kept in this study. The mandatory 4 were kept as well as those that directly correlated with the Sciences such as Physical Science (Physics and Chemistry), Life Sciences and Geography. The NBT marks were also used. The variables used in the final dataset are shown in Table I and are put into their respective categories according to the Tinto (1975) framework. HL stands for Home Language and FAL stands for First Additional Language. Lastly, the dataset was balanced so that each value of the target variable has an equal number of occurrences. The smallest class was the *QualMin* class with only 1069 records. The other two classes were then reduced to this amount by taking the most recent 1069 record i.e. undersampled.

D. Models

Six classification algorithms will be run in a similar fashion as Osmanbegovic *et al.* [16] and Ramesh *et al.* [15], which are used to predict the target variable of students. The six algorithms are: a Random Forest (RF), J48, Naive Bayes (NB), Logistic Regression (LR), Sequential Minimal Optimization (SMO) and a Multilayer Perceptron (MLP). The Naïve Bayes classification algorithm and the Logistic Regression will serve as baseline models. The MLP is a black box model, meaning that the structure of the network will not give any insights on the function being approximated.

The effectiveness of each model is tested through 10-fold cross validation. This is a re-sampling procedure by which a portion of the training data is not seen by the algorithm during training, but is used for validation. The dataset is first split randomly into a training and testing set. The training set is then split into K partitions (folds) where $K-1$ folds are used for training and the remaining fold is used for validation. This is then repeated until every K fold has served as the validation fold once. The validation accuracies are stored for every split. The model which gave the best validation accuracy is used and evaluated by the testing set. The test set is not included in the K-folds as this would bias the results and is therefore kept completely unseen.

The evaluation of the models comes by analysing the accuracy of the models predictions as well as the precision and recall values. Precision and recall are evaluation metrics calculated using the resulting confusion matrix. Figure 1 shows a confusion matrix and what each cell represents in a binary classification problem. The diagonal elements represents correctly classified instances. Precision, recall and accuracy are calculated as follows:

$$Precision = \left(\frac{TP}{TP+FP}\right),$$

$$Recall = \left(\frac{TP}{TP+FN}\right),$$

where TP are true positives, FP are false positives and FN are false negatives. The accuracy is calculated as

$$Accuracy = \left(\frac{TP+TN}{TP+FP+FN+TN}\right).$$

In our multiclass case, the confusion matrices will be 3×3 dimensional, so precision and recall will be calculated for each class. Both precision and recall are measures of accuracy of the model, but there exists a trade off between the two. Precision represents the proportion of results that are relevant while recall is the proportion of all relevant results that have been correctly classified. The trade off exists in that results will have to be repeatedly generated in order to recall everything thus lowering precision. An information gain ranking tool allows us to see which features played the biggest role in distinguishing which class a student falls into.

TABLE I
TABLE OF SELECTED FEATURES PLACED IN THEIR RESPECTIVE CATEGORY

Background	Individual	Pre-University
Home Province	NBTAL	Core Math
Age at First Year	NBTMA	Math Literacy
School Quintile	NBTQL	Additional Math
Rural or Urban	Year Started	English HL
International	Plan Description	English FAL
	Target/Outcome	Computer Studies
		Physical Sciences
		Life Sciences
		Geography
		Life Orientation

IV. RESULTS

Table II shows six models and their respective accuracies after 10-Fold cross validation. The Multilayer Perceptron took the longest to train while the other five models were

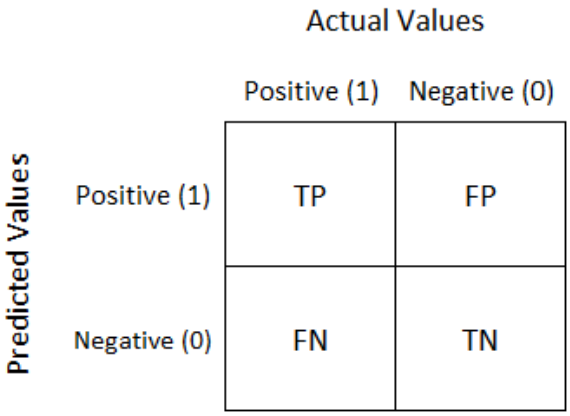


Fig. 1. Confusion matrix [17]

TABLE II
TABLE OF HIGHEST ACCURACY PER MODEL

Model	Accuracy
J48	**80.4%**
RF	79.6%
MLP	76.6%
NB	74.4%
SMO	71.9%
LR	70.0%

relatively fast. The J48 decision tree algorithm achieves the top accuracy at 80.4%. The RF model is not far behind. Our next best model, the MLP, jumps down by 3%. We can see that models that come from the decision tree paradigm (J48 and RF) outperform the black box and probabilistic models. Our NB baseline model sits third from the bottom. The SMO does worse than it and is therefore not a good approach in this scenario. The success of the J48 model is due to its features such as its ability to handle missing values, the continuous value range and threshold, that it does well in choosing, and its ability to prune the decision tree to remove branches that do not add value to the model [18]. This is consistent with Aly et al. [12] in showing the algorithm is robust and efficient. A property of trees that make it compatible with this data is that they work well with many features, especially categorical features. The worst performing model is our other baseline, the LR. The multivariate version uses the softmax activation function instead of a sigmoid activation function which is typically used for binary classification. Small differences in the data sometimes get blown out of proportion with the softmax function which may cause a bias to a certain class, hence the poorer performance.

Figures 2 to 7 depict confusion matrices for each model. In each case, it can be seen that the *Failed* value was the most accurately predicted class, *Qualifid* was second best and *QualMin* was the least accurately predicted class. Most confusion lies between the *Qualifid* and *QualMin* classes. The confusion between these classes comes from the fine line between the number of years it may take a student to graduate. If a student takes 3 years, they fall in the *QualMin* class, but if they take 4 they fall in the *Qualifid* class. The distinction between students in each class may not be clear from features such as Matric marks, NBT marks or even biographical profiles. Outside factors may cause a student to fail a year which is not quantified in our data and therefore can not be utilised to improve the predictive accuracy. However, the distinction between students who qualify overall and who fail may well be prevalent in features such as Matric marks and biographical profiles. This leads to the contrast in predictive accuracy between the values and reinforces the argument that all these features play a role in the success of a student's academic trajectory.

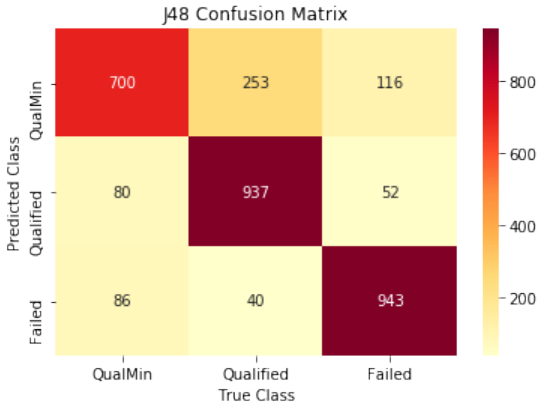


Fig. 2. Confusion matrix for the J48 model

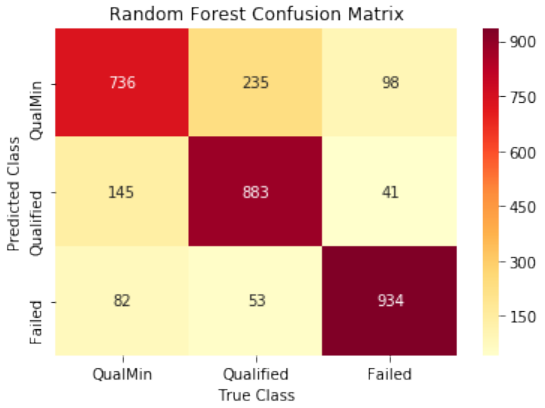


Fig. 3. Confusion matrix for the RF model

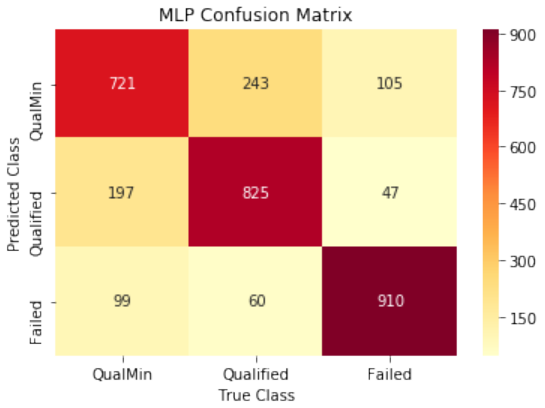


Fig. 4. Confusion matrix for the MLP model

Figures 9 to 11 are precision and recall graphs pertaining to the J48, Multilayer Perceptron and Logistic regression models respectively. The x-axis represents the recall while the y-axis represents the precision. In graph (c) of all three figures, we see the *Failed* value takes on a smoother curve than the others. As mentioned before, the *Failed* value has good accuracy amongst all the models with an average accuracy of 87.5%. We can see this performance is reinforced by the high recall each model produces. We are more interested in the recall

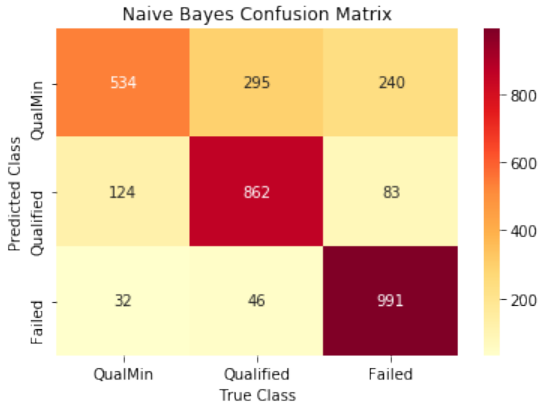


Fig. 5. Confusion matrix for the NB model

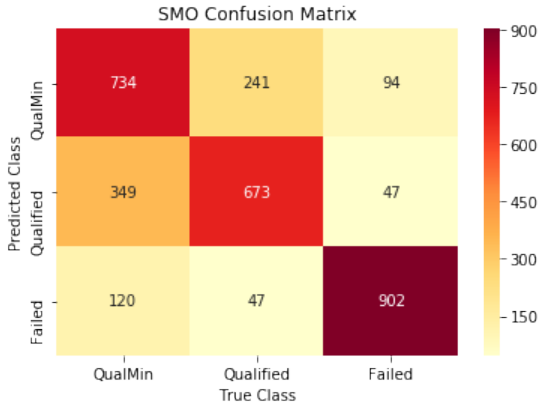


Fig. 6. Confusion matrix for the SMO model

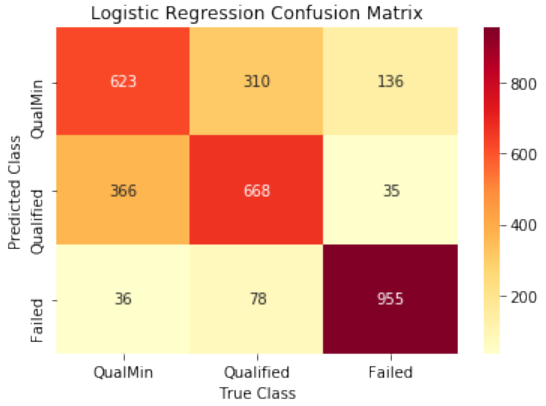
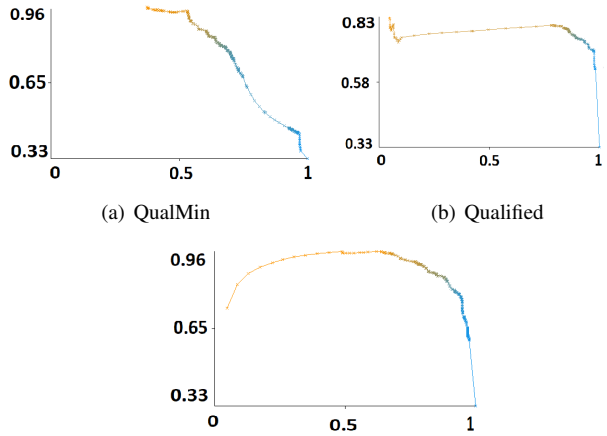


Fig. 7. Confusion matrix for the LR model

since it represents the proportion of relevant samples i.e. it represents how many students who failed which are identified by the model. The smoothness of the *Failed* curves in all three cases show that the model is not biased or over-fit. In all three cases we can see the *QualMin* suffers by the jaggedness of the curves. The models may be over-fitting at times here. Overall, the J48 model has the most amount of highest precision and recall values in Table III, but still has its flaws. Although not shown as a graph, the Naïve Bayes model had a very high recall for the *Failed* value, but did not match up with the other values.

TABLE III
PRECISION AND RECALL VALUES FOR EACH TARGET VALUE FOR EACH MODEL

	Precision			Recall		
	QualMin	Qualified	Failed	QualMin	Qualified	Failed
J48	0.808	0.762	0.849	0.655	0.877	0.882
RF	0.764	0.754	0.870	0.668	0.826	0.874
MLP	0.709	0.731	0.857	0.674	0.772	0.851
NB	0.774	0.717	0.754	0.500	0.806	0.927
SOM	0.610	0.700	0.865	0.687	0.630	0.844
LR	0.608	0.633	0.848	0.583	0.625	0.893

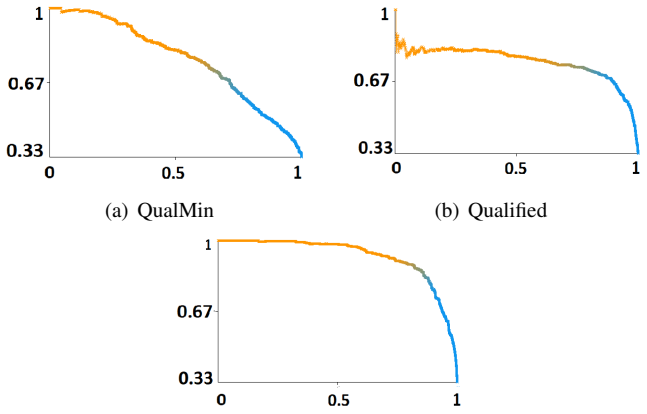


(a) QualMin (b) Qualified

(c) Failed

Fig. 8. Graphs showing precision and recall curves for each outcome class value for the J48 model

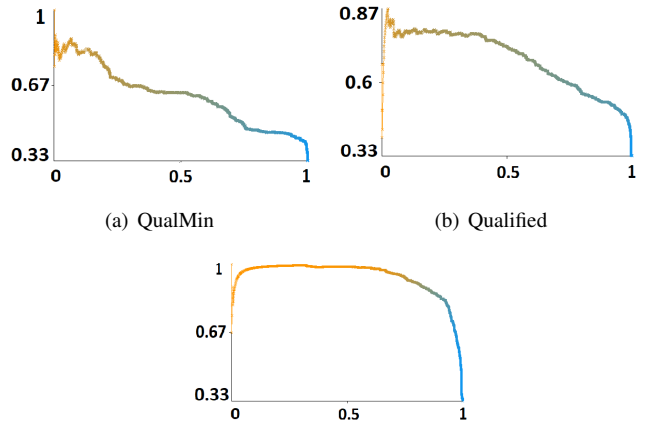
Table IV shows the features ranked according to their information gain. The first column indicates the ranking of the feature in terms of its contribution where the second columns indicates the feature's entropy value *(e)* where $0 \leq e \leq 1$. The colours indicate which of Tinto's categories the feature belongs to, corresponding to Table I. The top 7 most contributing features are: (1) The year the student is starting; (2) the programme the student is enrolling for; (3) the age of the student in their first year; (4) the high school quintile; (5) the NBT for academic literacy mark; (6) the NBT for quantitative literacy mark; and (7) the home province of the student. These results show the Background and Individual attribute groups of Tinto (1975) have a dominant role in predicting student's performance. This is consistent with the results found by Dr. Ajoodha [2]. 4 out of the 5 Individual features contained in the dataset fall within the top 7 contributing features which indicates that Individual features are the most deterministic features compared to the Background and Pre-University. With Pre-University features being the least contributing feature



(a) QualMin (b) Qualified

(c) Failed

Fig. 9. Graphs showing precision and recall curves for each outcome class value for the MLP model



(a) QualMin (b) Qualified

(c) Failed

Fig. 10. Graphs showing precision and recall curves for each outcome class value for the LR model

set, it brings into question why the APS, which we recall is made up of Pre-university subject marks only, is the primary mechanism for offering students places in programmes.

V. Conclusion

Due to the low pass rates of students in the Science faculty at a research intensive university in South Africa, there is a need for higher quality academic advisory so students may make better informed decisions about their academic trajectories. More specifically, a recommendation engine that will allow students to gauge their future success and academic affordability of a programme is proposed. In addition to this, the mechanism used to select students to be accepted to a programme, APS, is debated as being ill-informed. However, discovering the factors that influence a student's future success was not an easy task. Tinto (1975) outlines a framework that identifies three categories of student attributes that influence the attrition rate. In this research, we have employed this framework to identify which category of attributes has the biggest influence on Science students and compare this to the APS mechanism.

TABLE IV
Information Gain

Rank	Information Gain (e)	Feature
1	0.6696	Year Started
2	0.5685	Plan Description
3	0.3410	Age at First Year
4	0.0257	Quintile
5	0.0153	NBTAL
6	0.0118	NBTQL
7	0.0087	Home Province
8	0.0089	Life Sciences
9	0.0087	Physical Sciences
10	0.0081	Life Orientation
11	0.0076	Geography
12	0.0075	Core Mathematics
13	0.0073	NBTMA
14	0.0056	Rural/Urban
15	0.0049	English FAL
16	0.0032	English HL
17	0.0014	Additional Mathematics
18	0.0002	International
19	< 0.0002	Mathematics Literacy
20	< 0.0002	Computer Studies

Attributes were split into Background, Individual and Pre-University categories. Six classification models from different paradigms were run on the data to predict which of the following three classes a student falls into: QualMin (qualify in the minimum time of three years) Qualified and Failed. The models proved that the combination of these attributes can predict a student's outcome within a 70% - 80% accuracy. Decision tree algorithms, specifically the J48 algorithm, proved to be the most robust with high recall values and mostly the highest precision values as well as the highest accuracy overall. Furthermore, the most influential category of attributes was the Individual attributes which included the Plan description and NBT marks. The Background attributes which included the age at first year, school quintile and province of the student also proved to have high influence. Interestingly, the Pre-university attributes, which are high school grades per subject, had the least influence. APS is calculated as a sum of points pertaining to the marks received per high school subject. Thus we can see a mismatch in the requirements needed to earn a place in a programme and the likelihood of succeeding.

The overall contributions of this work is providing a

more complex way of viewing student placement in the Science faculty as opposed to using APS. Thus this research proposes a recommendation engine which takes in a combination of attributes of students' Background, Individual and Pre-university profiles and proposes programmes which will optimise their success and minimise time spent in doing so. This paper also provides insights into machine learning models that work well with educational data. We further show the importance of each attribute in the predictability of the models by ranking them according to their information gain as seen in Table IV.

The data used in this research is taken only from the 10 years between 2008 and 2018, which is limiting. Data from previous years may show other trends or influence the patterns differently. Another limitation is that the data is only from the Science Faculty.

Future work in this field may include extending the engine to accommodate all faculties in the university. This would greatly improve the output of the university as students would be in fields more suited to them or would know the effort they need to put in beforehand. Extending this to all universities in South Africa would show an improvement in the higher education sector. Another future implementation may be to look at first year students who are high risk, at the end of their first year, and propose a change of programme for them according to the models. As an improvement to the model, data relating to the Peer-Group Interactions and Faculty Interactions of students may be added and ranked as they are part of the framework of Tinto (1975) and may prove useful.

REFERENCES

[1] Ian Scott, Nan Yeld, and Jane 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] Ritesh Ajoodha, "Predicting learner attrition for the sciences using background, individual attributes, and schooling at a south african higher educational institute," *Private Communication*, 2019.

[3] Dina G Markowitz, "Evaluation of the long-term impact of a university high school summer science program on students' interest and perceived abilities in science," *Journal of Science Education and Technology*, vol. 13, no. 3, pp. 395–407, 2004.

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

[5] John P Bean, "Conceptual models of student attrition: How theory can help the institutional researcher," *New directions for institutional research*, vol. 1982, no. 36, pp. 17–33, 1982.

[6] Duan Van der Westhuizen and Glenda Barlow-Jones, "High school mathematics marks as an admission criterion for entry into programming courses at a south african university," *The Independent Journal of Teaching and Learning*, vol. 10, no. 1, pp. 37–50, 2015.

[7] Sarah Rauchas, Benjamin Rosman, George Konidaris, and Ian Sanders, "Language performance at high school and success in first year computer science," *SIGCSE Bull.*, vol. 38, no. 1, pp. 398–402, Mar. 2006.

[8] Pat Byrne and Gerry Lyons, "The effect of student attributes on success in programming," *ACM SIGCSE Bulletin*, vol. 33, no. 3, pp. 49–52, 2001.

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

[10] Ritesh Ajoodha and Ashwini Jadhav, "Identifying at-risk undergraduate students using biographical and enrollment observations for mathematical science degrees at a south african university," *Arctic Journal*, vol. 72, no. 7, pp. 42–71, 2019.

[11] Kamal Taha, "Automatic academic advisor," pp. 262–268, 2012.

[12] Walid Mohamed Aly, Osama Fathy Hegazy, and Heba Mohmmed Nagy Rashad, "Automated student advisory using machine learning," *International Journal of Computer Applications*, vol. 975, pp. 8887, 2013.

[13] Salvatore Ruggieri, "Efficient c4. 5 [classification algorithm]," *IEEE transactions on knowledge and data engineering*, vol. 14, no. 2, pp. 438–444, 2002.

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

[15] Vamanan Ramesh, P Parkavi, and K Ramar, "Predicting student performance: a statistical and data mining approach," *International journal of computer applications*, vol. 63, no. 8, pp. 35–39, 2013.

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

[17] Sarang Narkhede, "Understanding confusion matrix," date accessed: 06/09/19.

[18] Gaganjot Kaur and Amit Chhabra, "Improved j48 classification algorithm for the prediction of diabetes," *International Journal of Computer Applications*, vol. 98, no. 22, 2014.