# Using Student Characteristics to Promote Student Success at Higher-education Institutions

Jared Tremayne Naidoo[1], Ashwini Jadhav[2], Khanyisile Sixhaxa[3,*], and Ritesh Ajoodha[4]

[1] School of Computer Science and Applied Mathematics,
The University of the Witwatersrand, Johannesburg, South Africa
jared.naidoo1@students.wits.ac.za,
[2] Faculty of Science,
The University of the Witwatersrand, Johannesburg, South Africa
ashwini.jadhav@wits.ac.za ,
[3] School of Computer Science and Applied Mathematics,
The University of the Witwatersrand, Johannesburg, South Africa
khanyisile.sixhaxa1@students.wits.ac.za
[4] School of Computer Science and Applied Mathematics,
The University of the Witwatersrand, Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

**Abstract.** Accurately predicting students performance is useful in the higher education sector, and one key area of application would be the optimal placement of students within a university program. The main aim of this study is to propose an alternative approach into how students are admitted into universities as opposed to the traditional APS system for admission. To achieve this, we features like a student's background, individual and pre-college information and make use of six off the shelf machine learning classifiers, namely Random Forest, the K-Star Algorithm, Naive Bayes, a Multi-Layer Perceptron, the J-48 Decision Tree and Logistic Regression where we achieved accuracy of 95%, 93%, 78%, 91%, 88% and 92% respectively. These models can then be used to facilitate student placement. We also illustrate a relationship between background, individual and pre-college attributes of a student to academic performance in this paper. We makes use a feature extraction technique (Principle Component Analysis [15] to select the optimal number of features to feed into our 6 predictive models. The traditional method of aggregation of high school marks to deem whether a student has the ability to pass a specific curriculum is rejected. We argue that a more student centric system that looks at incorporating the students background, individual and pre-college attributes is needed in order to place the student in an academic program that will maximise the possibility of passing in minimum time.

**Keywords:** Machine Learning, Student Success, Student Characteristics

## 1 Introduction

In South Africa and many parts of the world, education is seen as a gateway to great employment opportunities and financial freedom. The authors in [17] reported that approximately one million students complete high school each year and approximately one fourth go on to university. This is a great opportunity for students, however 29% of first year students either fail or drop out or repeat first year of study [3]). This is due to several factors including academic performance, personal circumstances, pre-college education or misalignment with the academic curriculum.

This paper uses the conceptual framework of [24]. The [24] model of student attrition is regarded as one of the most credible student attrition conceptual frameworks found in modern literature. The author in [24] states that an individual's academic performance is a product of their Background, Individual and Pre- College characteristics and further adds that these three defining characteristics determine an individual's ability to set academic goals and stick to them. [24] also goes on to stress the importance of a good social life and involvement in university activities outside of the class to be a key success factor in students success.

The problem that we are challenged with however, is taking a student's background, individual, and pre-college attributes and providing the student with the optimal academic path for them to pursue at university. This can be achieved by gauging the "Risk Profile" of a student within a particular academic programme. We can then suggest or offer the student with the academic programme choices to minimise their "Risk Profile". A student is classified into three different "Risk Profiles". The risk profile is then used to assess whether a student will experience difficulty within a particular academic programme (Table 1).

Table 1: Description of the Risk Profile Assigned to a Student

| **Risk Profile** | **Risk Description** |
|---|---|
| Low Risk | Completed qualification in three years (min time) |
| Medium Risk | Completed qualification but took more than 3 years |
| High Risk | Failed to obtain a qualification |

A "Risk Description" is used to allocate a "Risk Profile" to a particular student (the process used to label training data). Various machine learning models are trained to identify a student's "Risk Profile" using their own background, individual and pre-college attributes based of historical data.

This study looks at students within the field of Earth Sciences at a research intensive university in South Africa. The field of Earth Science exhibits a unique characteristic- of the data collected, 45.5% of students within the Earth Sciences

pass the first year of university. This is considerably higher than the 29% found at the same university in another study [2].

This study uses data from 2010-2016 as this interval enables us to examine student attrition across a student's full 3 years towards a Bachelor of Science degree. It was noted that during 2010 - 2016, there were an estimated 1400 students that registered within the Earth Sciences; but only 770 entries were usable in this study (due to data quality issues such as missing values). The percentage of the students from 2010 to 2016 according to "Risk Profile" are as follows: Low Risk comprises of 16%, Medium Risk comprises of 61% and High Risk comprises of 23%.

It was noted that 84% of the students admitted to a course within the Earth Sciences do not complete the course in minimum time, although Earth Sciences graduated 77% of their student intake within this period. This highlights the importance of using the Earth Sciences data set for the study as we can infer to characteristics that make students successful within the field.

Currently the only means of gauging suitability of a student to an academic program is through the use of pre-college attributes, specifically high school academic performance. This includes the raw matric marks and a cumulative admission point score (APS). This metric does not factor in any of the background and individual attributes of the [24] framework, but just incorporates the pre-college academic performance input.

The authors in [4] noted that there are inaccuracies when using only the APS score as a means of a gauging student performance. It was found that there is an overlap between failing students and passing students with high APS scores. This proved that an additional metric is needed to gauge student performance.

The purpose of this study is to explore the relationship between background, individual and pre-college attributes to learner attrition, specifically we look to make use of this relationship as a means of advising students on the most suitable academic program for their own individual characteristics and personality. Additionally, the study uses the background, individual and pre-college attributes to classify students into low, medium and high risk profiles, which informs the proposed academic path

In the study, we trained 6 predictive machine learning models. The models trained include: a Random Forest, the K-Star Algorithm, Naive Bayes, a Multi-Layer Perceptron, the J-48 Decision Tree and a Logistic Regression. The models classify a student into one of the three "Risk Profiles" using the student's background, individual and pre-college attributes as features. We then evaluate the student's "Risk Profile" across multiple "plan codes" to evaluate their suitability.

A key attribute used in the study is the student "plan code". The plan code is the academic program that a student is enrolled for. A possible method of evaluating a student's performance in a specific plan would be to measure the student's "Risk Profile" across multiple plan codes. We then recommend a plan code to the student, where the student's risk profile was the lowest.

We have used confusion matrices to measure the performance of each of the six models and have applied Correlation Coefficient, Information Gain (Entropy)

and Principal Component Analysis (PCA) to the "Risk Profile" as a means of feature selection and extraction before modelling. It was found that a Random Forest classifier trained on three classes was the most accurate. The Random Forest obtained an accuracy of 95%, which was closely followed by the K-Star algorithm with an accuracy of 93% and then Logistic Regression at 92%.

This research contributes towards a greater understanding of student attrition in South Africa and provides a framework to study student attrition. The research also highlights an additional mechanism of informing students of their potential and risks within a particular academic program.

The remaining sections of the paper is structured as follows; Background and Related Work within the field of student attrition, the Methodology used for this study (Data Pre-Processing, Feature Selection and Extraction), the evaluation of the models and Conclusion.

## 2 Background and Related Work

Institutions of higher education in South Africa have faced quite a few challenges in recent years. These challenges include demographic and economical inequality amongst students and poor academic performance. With the increased unemployment rate in South Africa, the author in [17] has emphasised the importance of post high school qualifications and concluded that a post high school qualification is advantageous in finding a job. In subsection 2.1 we look into how the student attrition rate at universities is linked to a student's background, individual and pre-college attributes, followed by outlining the conceptual framework used in this study in subsection 2.2. In subsection 2.3 we do a comparison study of all the models used across different literature on student performance in higher education.

### 2.1 Link Between Student Attrition and Background, Individual and Pre-college Attributes

If you pick two students from a South African university at random, student A and student B, chances are that you will find that they both come from very different backgrounds. You may find that student A is supported emotionally and financially whilst student B may be supported emotionally but may not have the finances to pay for three years at university. This has an impact on student B's academic performance. The authors in [8] found that students who had families with absolutely no financial flexibility and who were in frequent financial crisis found it hard to be a university student, and it took a toll on their academic progress and the Condy in [11] found that a large number of students drop out for a range of reasons which include poor programme choice, social circumstances and financial reasons. Student performance is also influenced by their biographical associations and previous performance among other factors [3]. Campbell and McCabe in [7] look into the realm of students completing a degree based on their high school marks or SAT scores and they found that high

school rank (school quintile) or the quality of education taught at a high school level has a definite impact on the student's first year performance. Many of the characteristics such as family background (emotional and financial support), previous schooling, social circumstances and emotional state are all subsets of the three main attributes proposed by Tinto in [24] which are as follows: Background, Individual and Pre-College attributes.

### 2.2 Conceptual Framework

In this study we adapt the conceptual framework proposed by Tinto in [**?**]. The framework examines the link between Background, Individual and Pre-College attributes to student attrition.

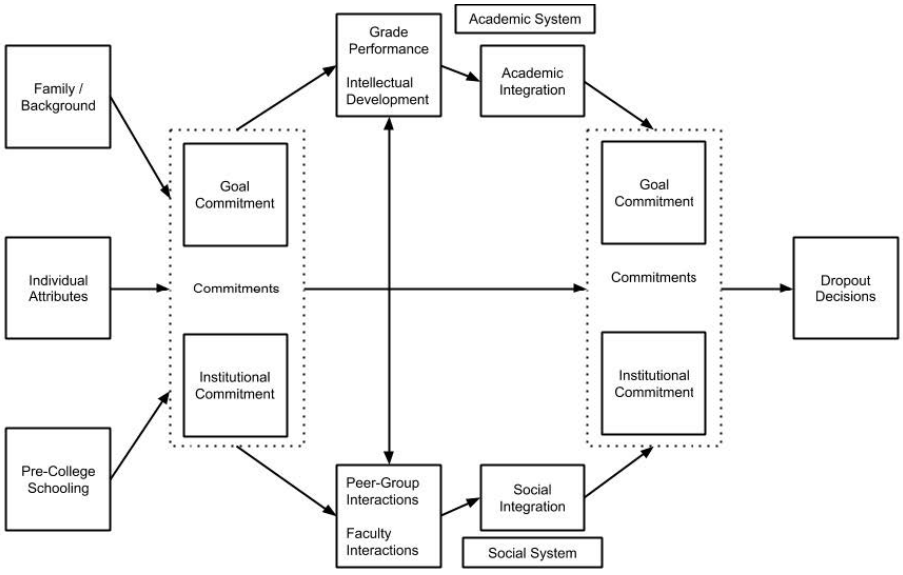



Fig. 1: Tinto [1975] model of student attrition which relates a student's Background, Individual and Pre-College attributes to success/failure at university.

The conceptual framework in Figure 1 contains three core input components: Family/Background, Individual and Pre-College Attributes. The framework speaks to the student's ability to set academic goals, their intellectual development and the ability to integrate into the social aspect of university. Tinto in [24] found that an improved goal commitment led to academic success and that social integration into the academic community produced strong academic results leading to a pass.

The three input attributes of the conceptual framework have proven to be good indicators of student attrition by multiple authors. Campbell and McCabe

in [7], utilised pre-college attributes i.e. the student's math, science and english SAT scores were used as means of predicting student performance. Ibrahim and Rusli in [14] performed a correlation coefficient analysis, and reported that a correlation of 0.87 on average between subjects, school characteristics and financial status in relation to success or failure at university.

Adams and Radix in [1] evaluated student's enrolment characteristics, they concluded that the level of success or failure could not be predicted using just enrolment information.

### 2.3 Comparison of Machine Learning Models used in Literature

In this section, we evaluate different machine learning models and techniques used to predict student attrition across different literature.

Decision Trees are one of the most used classifiers when dealing with inductive inference. It is a method of approximating discrete-valued functions. They are robust to noisy data and capable of learning disjunctive expressions. Decision trees classify instances by sorting them down the tree from the root node to some leaf node, which provides the classification of the instance [19]. The best performing model in this study is the Random Forest which is an ensemble of decision trees.

A Bayesian network can be described as a probabilistic graph model whereby the nodes represent random variables and the edges represent conditional independence assumptions. They provide a compact representation of joint probability distributions [2]. Bayesian networks are ideal as they can take an event that has occurred and calculate the probability of several known causes. We can then find the main contributing factor (i.e. what caused the outcome of a situation). The Naive Bayes model utilised in this study makes use of Bayesian Inference to deduce its results.

Thai-Nghe et al. in [20] compares the performance of a Decision Tree to Bayesian Network and they found that the performance of the decision tree was better than the bayesian network. We cannot assume that a decision tree would necessarily outperform a bayesian network, this is due to the vast difference in the dataset, number of observations and experiment configuration. A key take away from Thai-Nghe et al. in [20] is that the study contains 20 492 observations, of which 9765 observations are part of one class meaning that there is a large class imbalance. The dataset contained the classification labels of 'failed', 'fair', 'good' and 'very good'. This caused the model to potentially only classify students within the majority class with high accuracy.

Table 2 is a summary of the model performance examined in current literature. We have purposely selected literature that contained all 3 of the attributes in [24] to use as a measure of baseline performance.

After examining current literature, we can conclude that there is definitely a link between student attrition and the personal characteristics of a student. The papers [8,17,11,14,**?**,24] have provided strong evidence of the relationship and serve as good baselines for this research. In the following section, we will provide in-depth insight into the methodology utilised in this paper.

Table 2: Predictive Model Scores in Current Literature. Key: BG = Background; IN = Individual Attributes; and PC = Pre-college Grades

| Paper | BG | IN | PC | Model and Accuracy |
|---|---|---|---|---|
| [Osmanbegovic and Suljic 2012] | X | X | X | Naive Bayes - 76% |
| [Romero et al. 2012] | - | X | - | Decision Tree - 76% |
| [Osmanbegovic and Suljic 2012] | X | X | X | Neural Network - 71% |
| [Thai-Nghe et al. 2007] | X | X | X | Decision Tree - 86% |
| [Thai-Nghe et al. 2007] | X | X | X | Bayesian Netowork - 78% |

## 3 Methodology

In this study, we use background, individual and pre-college attributes to predict the "Risk Profile" of a student. There are three different "Risk Profiles" in this study. The first is Low Risk - completed qualification in 3 years (minimum time), Medium Risk - completed the degree in more than 3 years, and High Risk - failed to qualify for a degree. The study makes use of six different predictive machine learning models, which are as Random Forest, K-Star, Naive Bayes, Multi-Layer Perceptron, J-48 Algorith and Logistc Regression. The models vary in terms of architecture, configuration and training time. Confusion matrices have been provided to aid in the analysis of the models. A feature analysis providing the information gain and correlation towards each feature to the classification label ("Risk Profile") has been provided to gauge the strength of the features in relation to the "Risk Profile". In subsection 3.1 we outline the data collection and pre-processing step, followed by a brief overview of the feature selection and extraction process in 3.2. In the subsection 3.3, we predict at risk students and then evaluate the accuracy of our models and detail the ethics of this study in 3.4.

### 3.1 Data Collection and Pre-Processing

The data used in this study was collected at a South African University. The dataset contains background, individual and pre-college attributes of all students registered between the years 2010 - 2016 for a Earth Science degree. This period allows us to examine the complete three years of study towards a Bachelor of Science Degree. The dataset contains class imbalances whereby one of the classes contains significantly more observations than the other two classes. Nghe et. al. in [20] experienced an overfitting and a potential bias because of a similar set-up. Early trials in this study also indicated overfitting and a bias due to this class imbalance. This means that the predictive models could only predict one or two of the risk profiles accurately. To fix this and prevent overfitting, we employed a technique called Synthetic Minority Oversampling (SMOTE) which is used to remove class imbalance by synthetically generating observations within the minority classes [9].

### 3.2 Feature Selection and Extraction

Table 3 below highlights some of the key attributes in the study and allocates the characteristics to one of the three characteristics mentioned in the framework proposed by Tinto in [24] .

Table 3: Key attributes of a student (Found in our dataset) that are representative of Tinto's three characteristics

| **Background** | **Individual** | **Pre-College** |
|---|---|---|
| Home Country | Career Choice | Secondary School |
| Home Province | Year Started | - |
| School Quintile | Additional Language | - |
| Rural/Urban School | NBT | - |

Background includes the home country and province of the individual. An additional attribute used in the background category is the school quintile. Individual Attributes represent the individual's own capabilities. A strong attribute in this category is the plan code. We can relate the plan code to the student's future career choice and aspirations. The National Benchmark Tests (NBT) scores are used to gauge students proficiency in Mathematics and English. Pre-College Attributes include the student's performance in mathematics, english, science and a whole host of other subjects including geography, life orientation, economics, life sciences, civil and electrical technology. A full list of the attributes used in the experiments can be found in the evaluation section along with the Entropy and Correlation of each feature to the "Risk Profile".

Although we have access to 40+ features of a student, and even though many of these features potentially indicate the students "Risk Profile", the reality is that many of these features may not be of any value when seeking to predict the "Risk Profile". This presents us with the problem of feature selection [22]. Feature selection is the process of using a statistical or mathematical technique to gauge how important a specific attribute is in predicting the "Risk Profile" of a student. Information gain and Correlation were used as feature selection techniques. Information gain measures each feature's strength (Entropy) in relation to the label (Risk Factor). After performing feature selection and extraction on our dataset, reduced our features from 40 to 20 and ordered them in order of importance. Correlation measures the statistical relationship between two variables (i.e. if x increases,so does y) [12]. A full explanation and description of the feature selection techniques can be found in the evaluation section of this paper. Principle component analysis is a feature extraction technique used for dimensionality reduction. The idea here is to reduce the attributes from 40+ attributes to the lowest form (intrinsic dimension). We then find out which attributes con-

tributed the most in the data reduction. These attributes would be viewed as important and indicates that they are valuable attributes for modelling.

### 3.3 Prediction and Evaluation

We utilised aforementioned machine learning classifiers to make predictions on the label "Risk Factor". The modelling phase makes use of 10 Fold Cross Validation, which is a process whereby the data is split into k mutually exclusive subsets of equal size. The model is then trained k times on the k-folds of the data. The cross-validation estimate (accuracy) is the number of correct classifications divided by the number of instances in the data [16]. In the following paragraphs, we will provide a brief overview of each classifier.

*Random Forest* Random Forest is a combination of tree based predictors such that each tree depends on the values of a random vector sampled independently with the same distribution as all other trees in the forest [6]. An easy to understand description of a Random Forest is that it is an ensemble of multiple Decision Tree classifiers. Random forest are effective tools for classification problems. They rarely overfit due to the law of large numbers. Breiman in [6] states that Random Forest incorporates a certain essence of randomness when making a prediction which allows them to generalise well when classifying data.

*Naive Bayes* The Naive Bayes model (NBM) is the only model in this study that is based on Bayesian Statistics. The model is a simple and highly effective probability distribution algirithm [2]. It assumes all attributes of the examples are independent of each other, given the context of a class. This assumption is where the model gets the term "naive" from. In most real world tasks the NBM performs classification really well.

*Logistic Regression* A logistic regression is generally used in Binary Classification. However, it can be used in multi-class classification as well. Logistic regression solves these problems by applying the logit transformation to the dependent variable. In latent terms, the logistic model predicts the logit of Y from X [21].

*J.48 Decision Tree* The J.48 Decision tree is a popular tree based classifier. Originally known as the C4.5 classifier until it was renamed to the J.48 algorithm. The main function of this algorithm is to classify observations based on the entropy of multiple features to the classification label [5].

*K-Star* The K-Star algorithm first published in [10] was developed by Clearly and Trigg. The algorithm uses Information Theory to calculate the distance between two observations using a function called the K-Star function. It performs well in datasets that contain missing values [18].

*Multilayer Perceptron* The multilayer perceptron is a simple feed forward neural network that follows the basic feedforward process for prediction and back propagation for training (updating weights by back propagating a vector through the network [13]). The network used in this study utilises the basic sigmoid function as the activation function.

### 3.4 Ethics Clearance

This study has gone through a strict ethics clearance process and the data utilised has been annonymised. The ethics application for this study has been approved by the universities ethics department (Human Ethics Committee - Non Medical). The clearance certificate protocol number is: H19/08/28. In the next section, we discuss the results yielded from our research and provide an evaluation of the models used.

## 4 Experimental Results

In this section, we present the key findings of the study. The first section is feature selection and extraction, detailing the techniques and observations. The next section presents the findings of the modelling process, containing the confusion matrices for all 6 models, as well as, an analysis of the results obtained for each model.

### 4.1 Principle Component Analysis

Principle Component Analysis (PCA) is a feature extraction technique which produces a new set of features based on an original set of raw features. PCA is commonly used for dimensionality reduction, increasing interpretability and sometimes minimising information loss within data [15,23]. We are interested in the features used to build the resulting components (new features ) produced from PCA. The goal here is to perform PCA, i.e. Dimensionality reduction (from 41 features to 20 features) and observe which features are used the most by PCA to reduce the dimensionality of the data. This would indicate that much of the data's information or meaning lies in these features. Hence those features would be important to have in a feature set used for modelling. We aim to preserve 95% of the data's information (explained variance) when reducing components and, we are only interested in the top ranking features that allow us to preserve these principle components. We have included features that contributed to the top 3 principle (Plan Code, School Quintile and Home Language ) component's which preserve a 95% variance. The intrinsic dimensionality of the data used in this study is 7 dimensions. This is a significant reduction from 41 features down to 7 principle components (features).

The top 7 features include Plan Code, Plan Description, School Quintile, Home Province, Mathematics, Science and English Home Language, where the top three form part of [24] main attributes. This is a good indicator that the

features of [24] provide a strong underlying framework for predicting student attrition. Further more the attributes that are in the top three principle components are also featured as highly correlated features in Table 5.

### 4.2 Model Results and Evaluation

In this section, we produce the results of the 6 predictive models in the form of 6 confusion matrices followed by an analysis of each of the six models. The models used in this study are as follows: a Random Forest, the K-Star Algorithm, Naive Bayes, a Multi-Layer Perceptron, the J-48 Decision Tree and a Logistic Regression. The 6 models makes use of 10 fold cross validation

*Random Forest Evaluation* The Random Forest was the best performing model. It scored 731 out of 768 labels correctly. The misclassifications are fairly distributed which indicates that this model generalises well. The model did not perform as well with the "Medium Risk" category compared to the other 2 labels, however, it's classification of "Medium Risk" is still better than the other models. This model was very quick to train and has obtained an overall accuracy of 95%.

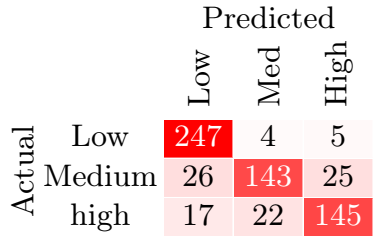
*K-Star Evaluation* The K-Star model was the second best performing model. It scored 719 out of 768 labels correctly. The model performed the best when classifying "High Risk" labels and performed the worst on "Low Risk" labels. The model also classified "High Risk" more accurately than the Random Forest (RF misclassified 9; this misclassified 8). Overall this models performance is incredible given that it uses the distance between points as metric of classification verses a more statistical approach. This model was very quick to train and has obtained an overall accuracy of 93%.

*Naive Bayes Evaluation* The Naive Bayes model was one of the poorer performing models, correctly classifying 592 out of 768 labels. It performed particular badly in classifying "High Risk" labels by misclassifying a total of 74 labels. The model was quick to train and has obtained an overall accuracy of 78%.

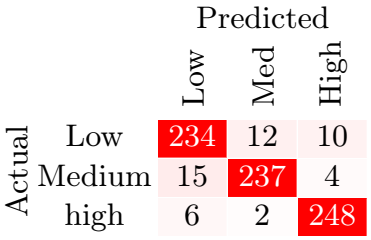
*Multilayer Perceptron Evaluation* The Multilayer Perceptron is the only model in the study that uses a "black box" technique. It took the longest time to train. It was the fourth best performing model and correctly classified 699 out of 768 labels. The misclassifications are fairly similar to the Random Forest and K-Star models which, like the other two models, tells us that this model generalises well. This model scored an overall accuracy of 91%.

*J-48 Decision Tree Evaluation* The J-48 decision tree scored an overall average of 88% and correctly classified 674 out of 768 labels. The model did not perform as well with the "Low Risk" category compared to the other two categories. This model was very quick to train and has obtained an overall accuracy of 95%.

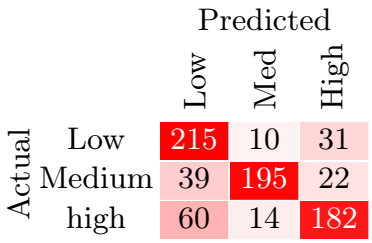
*Logistic Regression Evaluation* The Logistic Regression was the third best performing model. It scored 707 out of 768 labels correctly. The model's performance is very similar to the Random Forest and K-Star models. This model was very quick to train and has obtained an overall accuracy of 92%.



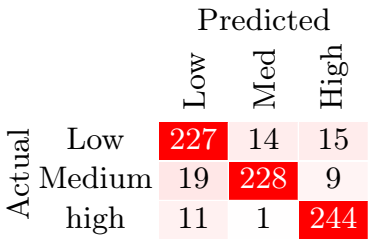

(a) Random Forest - 95% Accuracy, Correctly identified 731/762 labels



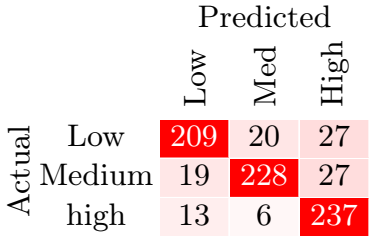

(b) K-Star - 93% Accuracy, Correctly identified 719/762 labels



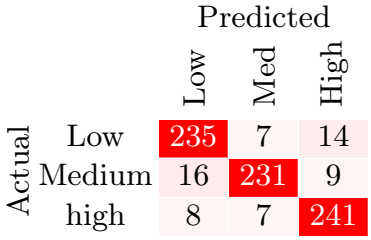

(c) Naive Bayes - 78% Accuracy, Correctly identified 592/762 labels




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




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




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

Fig. 2: 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.

### 4.3 Model Results Summary

The three best performing models were the Random Forest followed by the K-Star model and then the Logistic Regression model. The Random Forest correctly classified 95% of the data whilst the other two were lower with averages of 93% and 92% respectively. The Random Forest generalises well and the distribution of scores for misclassified cases are evenly distributed across the three classes (Low, Medium and High Risk). It must be noted that model performance greatly increased once I evenly balanced the classes (i.e. "Risk Profile"). Previously, model accuracies would not exceed 65% with a class imbalance present. In terms of the Misclassifying Risk (cases when we incorrectly classify a student), the top three models contain much of the misclassification risk across the medium risk class verses the low and high risk classes. This may be an indication that the student could still take the academic path but with an understanding that he/she may experience difficulty along the way. The university can then provide the student with assistance (taking a preemptive approach). A method of improving the models' scores could be to ensemble the top three models by using a fourth classifier. This fourth classifier would take the three models as input features. It would then learn which models predict particular classes better than others. This may give us a better overall average and prediction in the end. To conclude, the results obtained are very good and prove that the features in [24] - Background, Individual and Pre-College attributes can provide accurate indications of an individuals risk within a particular academic program.

### 4.4 Implications, Conclusion and Future Work

Given the current situation of Higher Education in South Africa being the limited spaces available and the fact that university intake increase on a yearly basis. South African universities need to come up with a mechanism to deal with the large volume of students coming in to university. One of the ways that this can be achieved is to produce more graduates in minimum time. This can only be achieved if students pass all of their modules in minimum time. Universities need to help students increase their chances of passing all of the modules selected. A system like the one featured in this study (a course recommender system), or a system that can detect students experiencing difficulty, can help universities achieve a pro-active approach towards reducing student attrition. This research provides evidence that a "course recommender" system which uses a student's Background, Individual and Pre-College attributes as a means to predict student risk for a particular academic path is possible. It can produce reliable results that can have an impact on the academic path a student takes as well as potentially reduce their duration of study. The limitations faced in this study was that we used a small data set and thus limited data instances to train our models, this slightly increases the bias of our models resulting in reduced accuracy. Overall, the models predicted each of the three classes fairly evenly and did not favour or bias one class over the other. A suggested future improvement would be to ensemble the top three classifiers so as to produce one super model that would

take input from the three top performing models and produce one result. This paper serves as proof that Background, Individual and Pre-College attributes can be used to predict student attrition which can then be used to recommend courses to students. It contributes towards a greater understanding of student attrition and highlights difficulties that many experience in South Africa. The recommender system would allow universities to take pre-emptive measures on student intake and can be used to help existing students during the year. To conclude, this study examined student attributes and characteristics that are normally overlooked. The study proves that background, individual and pre-college attributes should definitely be considered when admitting students to a university program.

## Acknowledgement

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

## References

1. Adams, R.V., Radix, C.A.: Predicting student performance in a caribbean engineering undergraduate programme. West Indian Journal of Engineering **41**(1) (2018)
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. Ajoodha, R., Jadhav, A., Dukhan, S.: Forecasting learner attrition for student success at a south african university. In: Conference of the South African Institute of Computer Scientists and Information Technologists 2020. pp. 19–28 (2020)
4. Ajoodha, R., Rosman, B.: Tracking influence between naïve bayes models using score-based structure learning. In: 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech). pp. 122–127. IEEE (2017)
5. Ali, M.M., Qaseem, M.S., Rajamani, L., Govardhan, A.: Extracting useful rules through improved decision tree induction using information entropy. arXiv preprint arXiv:1302.2436 (2013)
6. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
7. 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)
8. Case, J.M., Marshall, D., McKenna, S., Disaapele, M.: Going to university: The influence of higher education on the lives of young South Africans. African Minds (2018)
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
10. Cleary, J.G., Trigg, L.E.: K*: An instance-based learner using an entropic distance measure. In: Machine Learning Proceedings 1995, pp. 108–114. Elsevier (1995)

11. Condy, J.: Telling stories differently: Engaging 21st century students through digital storytelling. AFRICAN SUN MeDIA (2015)
12. Ezekiel, M.: Methods of correlation analysis. (1930)
13. Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: Applied logistic regression, vol. 398. John Wiley & Sons (2013)
14. Ibrahim, Z., Rusli, D.: Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. In: 21st Annual SAS Malaysia Forum, 5th September (2007)
15. Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **374**(2065), 20150202 (2016)
16. Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai. vol. 14, pp. 1137–1145. Montreal, Canada (1995)
17. Marock, C.: Grappling with youth employability in south africa (2008)
18. Martínez-López, Y., Madera-Quintana, J., De Varona, I.L.: Study of the performance of the k* algorithm in international databases. Revista Politécnica ISSN **2351** (1900)
19. Mitchell, T.M., et al.: Machine learning (1997)
20. Nghe, N.T., Janecek, P., Haddawy, P.: A comparative analysis of techniques for predicting academic performance. In: 2007 37th annual frontiers in education conference-global engineering: knowledge without borders, opportunities without passports. pp. T2G–7. IEEE (2007)
21. Peng, C.Y.J., Lee, K.L., Ingersoll, G.M.: An introduction to logistic regression analysis and reporting. The journal of educational research **96**(1), 3–14 (2002)
22. Ramaswami, M., Bhaskaran, R.: A study on feature selection techniques in educational data mining. arXiv preprint arXiv:0912.3924 (2009)
23. Reddy, G.T., Reddy, M.P.K., Lakshmanna, K., Kaluri, R., Rajput, D.S., Srivastava, G., Baker, T.: Analysis of dimensionality reduction techniques on big data. IEEE Access **8**, 54776–54788 (2020). https://doi.org/10.1109/ACCESS.2020.2980942
24. Tinto, V.: Dropout from higher education: A theoretical synthesis of recent research review of educational research, 45 (1), 89-125. Retrieved from JSTOR website: http://www. jstor. org/stable/1170024 (1975)