University of the Witwatersrand

WITS
UNIVERSITY

School of Computer Science and Applied Mathematics
Faculty of Science
Honours Research Report

Using Background, Individual and Pre-College Attributes for Student Placement in the Earth Sciences

Jared Tremayne Naidoo

Student Number: 719238

Supervisors: 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, Jared Tremayne Naidoo, (Student number: 719238) am a student registered for COMS4059A 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: 

Signed on the **20** day of **October , 2019** in Johannesburg.

Abstract

Accurately predicting student performance is useful in multiple applications. One key area would be the optimal placement of students within a university program. In the study by Ajoodha and Jadhav [2019] they found that almost 29% of first year students either repeat, or drop out at a Research Intensive University in South Africa. Given the large number of failing first-year students, a solution to this problem is urgently needed.

In this study I use the model of student attrition found by Tinto [1975] to accurately predict student performance in a particular program. The key attributes of a student used in the study are background, individual and pre-college information. This report includes an in-depth analysis of the performance of these attributes when used to train 6 different predictive machine learning models. These models can then be used to facilitate student placement. The study goes on to illustrate a relationship between background, individual and pre-college attributes of a student to academic performance.

This study makes use of two feature selection techniques (Correlation Coefficient - [Ezekiel 1930] and Information Gain - [Ajoodha 2019]) and one feature extraction technique (Principle Component Analysis - [Jolliffe and Cadima 2016]), which allowed the researcher to select the optimal number of features (attributes) to feed into the 6 predictive models. The predictive models used in this study are: a Random Forest, the K-Star Algorithm, Naive Bayes, a Multi-Layer Perceptron, the J-48 Decision Tree and lastly a Logistic Regression. An in-depth analysis of the results and confusion matrices for each of the 6 models is included in the results section. The analysis of the feature selection/extraction techniques showed three different techniques producing the same similar set of features which are subsets of the Tinto [1975] features (Background, Individual and Pre-College). The confusion matrices indicates the accuracy of 6 different models when predicting student performance. The analysis of the confusion matrices highlights the pros and cons of each model.

The traditional method of aggregation of high school marks to deem whether a student has the ability to pass a specific curriculum is rejected. I 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. This ensures that a student matches a program using multiple facets of student information.

Contents

Introduction

It is well known that through education you can achieve anything in life. In South Africa and many parts of the world, education is seen as a gateway to great employment opportunities and financial freedom.

In Marock [2008] they found 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 many fail the first year and either drop out or repeat (29% of first year students repeat or drop out Ajoodha and Jadhav [2019]). This can be the result of multiple factors including academic performance, personal circumstances, pre-college education or misalignment with the academic curriculum.

This study uses the conceptual framework of Tinto [1975]. The Tinto [1975] model of student attrition is regarded as one of the most credible student attrition conceptual frameworks found in modern literature. Tinto [1975] states that an individual's academic performance is a product of their Background, Individual and Pre-College characteristics. Tinto further went on to say that these three defining characteristics determine an individual's ability to set academic goals and stick to them. Tinto [1975] 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 as well.

The problem that we are challenged with 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 provide the student with the academic programme with which their "Risk Profile" is the lowest.

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.

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

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

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 new student's "Risk Profile" using their own background, individual and pre-college attributes based of historical data.

The 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 [Ajoodha and Jadhav 2019].

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 found that during 2010 - 2016 there were an estimated 1400 students that registered within the Earth Sciences; only 770 were usable in this study (due to data quality issues - 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%. Figure 1 is a plot of the "Risk Profile" to students over the specified interval.

This proves that there is a serious problem whereby 84% of the students admitted to a course within the Earth Sciences do not complete the course in minimum time. What is important to note is that the Earth Sciences graduated 77% of their student intake within this period. This highlights the importance of using students within Earth Sciences for the study as we can infer characteristics that make students successful within the field.

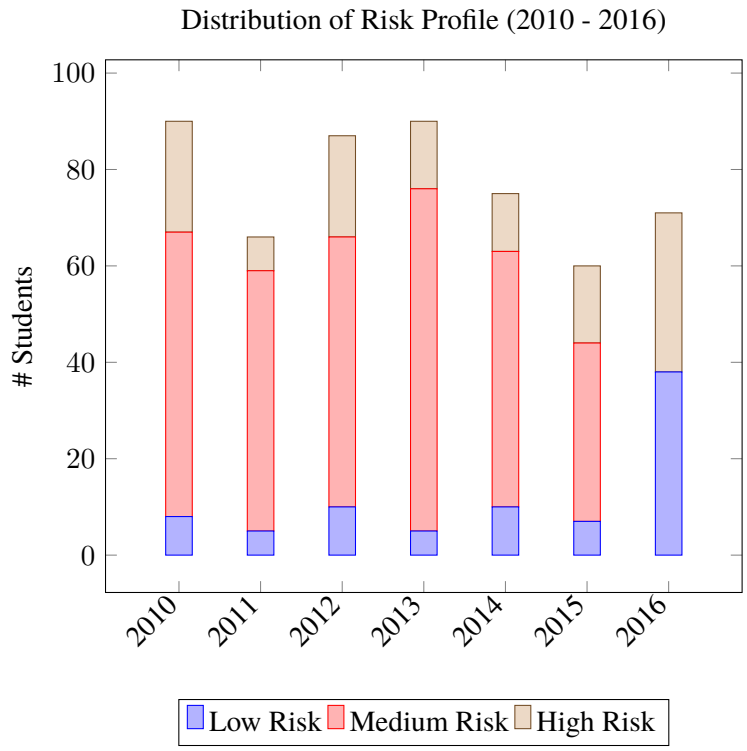

Figure 1: Distribution of Risk Profile between 2010 and 2016. Displaying the percentage of Low, Medium and High Risk profiles within the Faculty of Science, specifically students studying towards a degree in the field of Earth Sciences.

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 marks and a cumulative admission point score (APS). This metric does not factor any of the background and individual attributes of the Tinto [1975] framework, rather just incorporating the pre-college academic performance.

In Ajoodha [2019] he found that there are in-accuracies 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 aim of this study is to use additional student information (Background and Individual attributes) to gauge a student's suitability for an academic program rather than simply using academic performance of their pre-college schooling. A student's suitability would be gauged according to their "Risk Profile" for a specific program. This would provide a pro-active approach of advising students on suitable academic paths for their unique biographical profile. This minimises the possibility of the student failing and maximises the chance of passing in the minimum time.

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. The study uses the background, individual and pre-college attributes to classify students into low, medium and high risk profiles. These characteristics as depicted in Tinto [1975] form the basis of this study.

In the study I 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.

I have elected to use confusion matrices to measure the performance of each of the six models and additionally elected to use the 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%. It was closely followed by the K-Star algorithm with an accuracy of 93% and the Logistic Regression at 92%.

This research contributes towards a greater understanding of student attrition in South Africa, it provides a framework to study student attrition and acknowledges many of the challenges faced by students. The research highlights an additional mechanism of informing students of their potential 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 a conclusion.

Background and Related Work

Institutions of higher education in South Africa have faced quite a few challenges in recent years. These challenges include inequality amongst students, poor academic performance, and multiple student protests. In 2018 South Africa had an all time high unemployment rate of 27.5% [Smit 2018]. Marock [2008] has emphasised the importance of post high school qualifications when searching for employment. They concluded that a post high school qualification is advantageous.

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. In Jennifer M. Case and Mogashana [2018] they 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.

In Chetty [2015] it was found that a large number of students drop out for a range of reasons which include poor programme choice, social circumstances and financial reasons. A student's social circumstances, financial and environmental factors all have an impact on the student's performance at university. Student performance is influenced by their biographical associations and previous performance among other external factors [Ajoodha and Jadhav 2019].

The paper Campbell and P. McCabe [1984] looks into the realm of students completing a degree based on their high school marks or SAT scores. 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 Tinto attributes Tinto [1975] - Background, Individual and Pre-College attributes.

Conceptual Framework

In this study I adopt the conceptual framework found in Tinto [1975]. The framework examines the link between Background, Individual and Pre-College attributes to student attrition. Figure 2 depicts the well known Tinto Framework found in Tinto [1975].

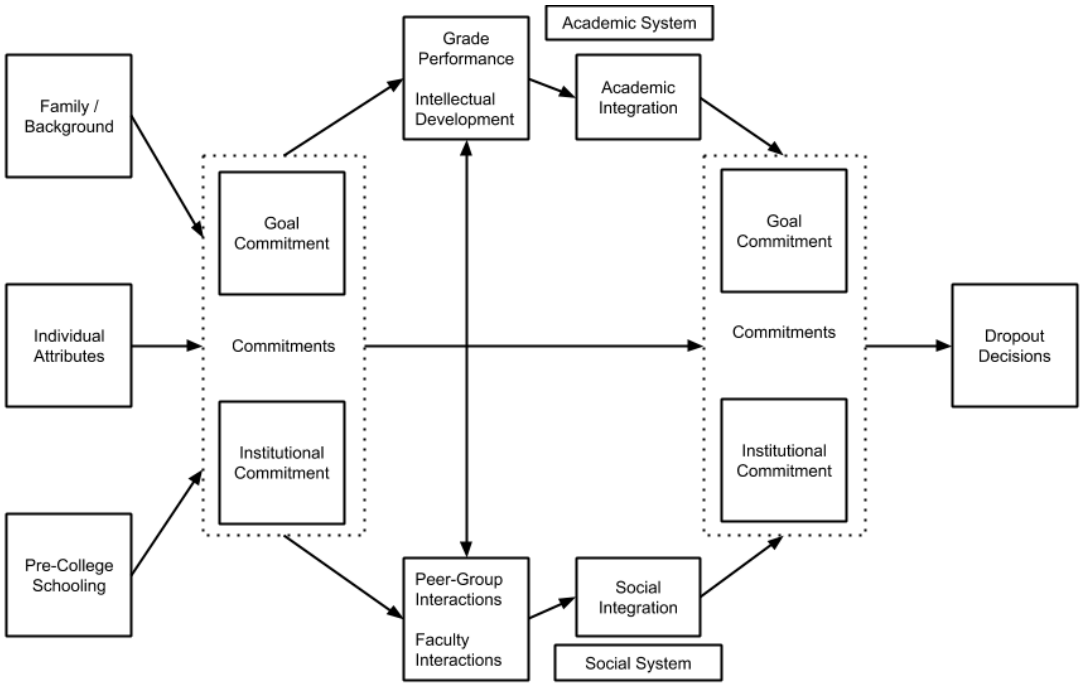

Figure 2: Tinto [1975] model of student attrition which relates a student's Background, Individual and Pre-College attributes to success/failure at university

The conceptual framework in Figure 2 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 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 [Tinto 1975].

The three input attributes of the conceptual framework have proven to be good indicators of student attrition by multiple authors. In Campbell and P. McCabe [1984], they utilised pre-college attributes to predict student performance. i.e. The student's math, science and english SAT scores were used as means of predicting student performance. In Ibrahim and Rusli [2007] they performed a correlation coefficient analysis. They found a correlation of 0.87 on average between subjects, school characteristics and financial status in relation to success or failure at university.

In the paper Adams and Radix [2018] they evaluate a student's enrolment characteristics. The goal was to predict the final grade and in course performance using just pre-college data. They found that enrolment information can be used to predict success or failure, however, they concluded that the level of success or failure could not be predicted using just enrolment information.

Comparison of Machine Learning Models used in Literature

In this section I evaluate different machine learning models and techniques used to predict student attrition in current literature.

Decision Trees are one of the most used techniques 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 [Mitchell 1997]. The best performing model in this study is the Random Forest [Breiman 2001] 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 [Ajoodha 2018]. 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 the situation? The Naive Bayes model utilised in this study makes use of Bayesian Inference to deduce its results.

Thai-Nghe et al. [2007] compares the performance of a Decision Tree to Bayesian Network. The study 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. [2007] is that the study contains 20 492 observations, of which 9765 observations are part of one class. i.e. This study contains 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. I have purposely selected literature that contained all 3 of the Tinto [1975] attributes to use as a measure of baseline performance.

Paper	**Background**	**Individual**	**Pre-College**	**Model & Accuracy**
[Osmanbegović and Suljic 2012]	Included	Included	Included	Naive Bayes - 76%
[Romero et al. 2012]	-	Included	-	Decision Tree - 76%
[Osmanbegović and Suljic 2012]	Included	Included	Included	Neural Network - 71%
[Thai-Nghe et al. 2007]	Included	Included	Included	Decision Tree - 86%
[Thai-Nghe et al. 2007]	Included	Included	Included	Bayesian Netowork - 78%

Table 2: Predictive Model Scores in Current Literature

Background and Related Work Summary

The section began with an analysis of student attrition in current literature. The sub-sections started by examining a potential link of student attrition to background, individual and pre-college attributes. After examining current literature I can conclude that there is definitely a link between student attrition and the personal characteristics of a student. The papers Jennifer M. Case and Mogashana [2018], Marock [2008], Chetty [2015], Ibrahim and Rusli [2007], [Ajoodha and Jadhav 2019] and Tinto [1975] have provided strong evidence of the relationship and serve as a good baselines for this research.

The remaining section discussed the performance of various predictive models used to predict student attrition in current literature. The purpose of the sub-section is to provide a baseline idea of the predictive power of machine learning within the context of student attrition. Four different papers indicated varied results which, going forward, can be used as baseline comparisons for the results found in this study.

Research Methodology

In this study I 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. 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".

This methodology section contains the following subsections; Data Collection and Pre-Processing, Feature Selection and Model Discussion.

Data Collection and Pre-Processing

The data used in this study was collected by the Academic and Information Systems Unit (AISU) at a South African Research Intensive University. The AISU has granted me, the researcher, permission to use this data to carry out the experiments in the study. The dataset contains background, individual and pre-college attributes of all students registered between the years 2010 - 2016 within the Earth Sciences department of the university. 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. In [Thai-Nghe et al. 2007] they experienced 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/bias, I employed a technique called Synthetic Minority Oversampling (SMOTE - [Chawla et al. 2002]) which is used to remove class imbalance by synthetically generating observations within the minority classes. The final count of the risk profile labels used in the study are found in Table 3.

Risk Profile	Original Count	Experiment Count
Low Risk	107	254
Medium Risk	235	254
High Risk	428	254

Table 3: Distribution of "Risk Profile" after using "SMOTE" to remove class imbalances

Feature Selection and Extraction

According to Tinto [1975] an individual's performance is greatly influenced by three factors: 1) Background Attributes - includes the individual's family, 2) Individual Attributes and 3) The individual's Pre-College attributes. These three terms broadly represent numerous characteristics of the individual. Table 4 below high-

lights some of the key attributes in the study and allocates the characteristics to one of the three characteristics mentioned in [Tinto 1975].

Background	**Individual**	**Pre-College**
Home Country	Career Choice (Plan Code)	High school marks for a host of courses
Home Province	Year Started	-
School Quintile	Additional Language	-
Rural/Urban School	National Benchmark Tests (NBT)	-

Table 4: Key attributes of a student (Found in our dataset) that are representative of Tinto's three characteristics

Background includes the home country and province of the individual. An additional attribute used in the background category is the school quintile. The idea is that together these three features can tell us more about where the student comes from in terms of their background in South Africa.

Individual Attributes represent the individual's own capabilities. A strong attribute in this category is that of 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".

Techniques Used

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 [Ramaswami and Bhaskaran 2009]. 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). The result is the selection of the top 20 features that are ranked the highest.

Correlation measures the statistical relationship between two variables [Ezekiel 1930] (i.e. if x increases, does y increase as well?). 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 contributed the most in the data reduction. These attributes would be viewed as important and indicates that they are valuable attributes for modelling.

Prediction and Evaluation

This study utilises 6 different predictive modelling techniques. The following six algorithms were used to make predictions on the label "Risk Factor". The modelling phase makes use of 10 Fold Cross Validation. K-Fold Cross Validation 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 diveded by the number of instances in the data [Kohavi et al. 1995].

Random Forest

A 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 [Breiman 2001]. 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 2001] and incorporate 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 [Ajoodha 2019]. 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 [Peng et al. 2002].

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 [Ali et al. 2013].

K-Star

The $K*$ or K-Star algorithm first published in Cleary and Trigg [1995] was developed by John G.Clearly and Leonard E. Trigg. The algorithm uses Information Theory to calculate the distance between two observations using a function called the $K*$ function. It performs well in datasets that contain missing values [Martínez López et al. 2016].

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 [Hosmer Jr et al. 2013]). The network used in this study utilises the basic sigmoid function as the activation function.

Ethics Clearance

This study has gone through a strict ethics clearance process. The study contains observations of learners who is/ are/ had studied at a Research Intensive University in South Africa. The ethics application for this study has been approved by the universities ethics department (Human Ethics Committee - Non Medical). Strict methods have been put in place to protect the identity of the participants (anonymised data). The clearance certificate protocol number is: H19/08/28.

Experiments

In this section I 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.

Feature Selection

In this section I aim to explore the features (attributes) of the raw dataset (41 attributes). I aim to select a subset of these features to use in the modelling portion of the study. The feature selection techniques carried out include a correlation and information gain analysis. A feature extraction technique called Principle Component Analysis is used to extract meaningful features.

Correlation Analysis

A correlation analysis is usually performed between two or more variables to measure the relationship between these variables [Ezekiel 1930]. Table 5 below contains the strongest features in terms of correlation to the "Risk Profile" feature. The Feature Rank, Tinto [1975] category, Feature name and the Correlation value are included as well.

Ranking	Tinto Category	Feature	Correlation
1	Individual	PlanDescription	0.2011
2	Individual	PlanCode	0.1852
3	Individual	Majors	0.1537
4	Individual	YearStarted	0.1131
5	Individual	RegistrationStart	0.1131
6	Individual	AgeatDropOutORGrad	0.1074
7	Pre-College	AdditionalMathematics	0.1069
8	Pre-College	Electrical	0.0979
9	Individual	RegistrationEnd	0.0951
10	Pre-College	BusinessEconomics	0.0783
11	Pre-College	ComputerStudies	0.0777
12	Pre-College	LifeSciences	0.0744
13	Pre-College	MathematicsMatricMajor	0.0727
14	Individual	AgeAtFirstYear	0.0707
15	Pre-College	LifeOrientation	0.0689
16	Pre-College	CraftSpeechDrama*	0.0682
17	Background	SchoolQuintile	0.0667
18	Pre-College	Geography	0.0655
19	Individual	NBTQL	0.0578
20	Pre-College	PhysicsChem*	0.0571

Table 5: Top 20 strongest features (Ranked by Correlation to "Risk Profile")

Table 5 indicates that the top features of the correlation analysis are Plan Code, Plan Description, Majors, Year Started and more. We cannot assume that these are the best features to use in the model based on this first analysis. We perform two more techniques to evaluate feature strength from different perspectives using different metrics.

Feature Importance using Information Gain

This section explores the 41 features in terms of the Information Gain (IG) to the "Risk Factor". Table 6 highlights the ranking of the feature (1 - Highest IG, 20 Lowest IG), Tinto category, feature name and entropy score.

Feature Ranking	Tinto Category	Feature	Entropy
1	Individual	PlanCode	0.4612097
2	Individual	PlanDescription	0.4467982
3	Individual	YearStarted	0.3576305
4	Individual	RegistrationStart	0.3576305
5	Individual	Majors	0.2591897
6	Individual	AgeatDropOutORGrad	0.2295773
7	Background	RegistrationEnd	0.1774866
8	Background	AgeAtFirstYear	0.0876897
9	Background	Homeprovince	0.0572927
10	Pre-College	MathematicsMatricMajor	0.0554799
11	Individual	NBTQL	0.0387015
12	Pre-College	LifeOrientation	0.0373123
13	Background	SchoolQuintile	0.033505
14	Pre-College	PhysicsChem	0.0306236
15	Individual	NBTMA	0.0302447
16	Pre-College	AdditionalLanguage	0.0299495
17	Pre-College	Geography	0.0265736
18	Pre-College	LifeSciences	0.0221204
19	Individual	NBTAL	0.0192506
20	Pre-College	EnglishFirstAdditional	0.0119203

Table 6: Top 20 Features (Ranked by Entropy)

Feature Rank Plot

Figure 3 below contains the top 20 features from the Information Gain analysis plotted between Feature Rank and Entropy approaches. What we see is that from a rank of 10 onwards the importance of the feature in terms of Entropy is approaching 0. This tells us that these features are less important.

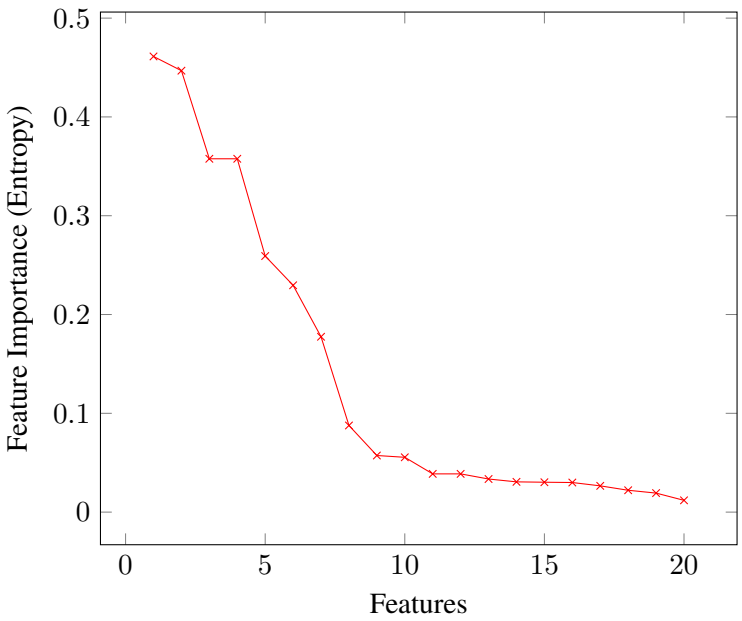

Figure 3: Plot of Feature Importance - Top 20 features (Ranked by Entropy)

Principle Component Analysis

Principle Component Analysis (PCA) is a feature extraction techniques 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 [Jolliffe and Cadima 2016].

I am 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

I aim to preserve 95% of the data's information (explained variance) when reducing components and, I am only interested in the top ranking features that allow us to preserve these principle components.

Table 7 below illustrates the ranking of the top features which contributed to the PCA dimension reduction. I have included features that contributed to the top 3 principle 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).

Component	Variance	Tinto Category	Feature
1	0.33	Pre-College	Life Sciences
1	0.33	Pre-College	Mathematics
1	0.33	Pre-College	Physical Sciences
1	0.33	Individual	NBTQL
1	0.33	Individual	Year Started
1	0.33	Background	School Quintile
1	0.33	Individual	Plan Code
1	0.33	Individual	Majors
1	0.33	Background	Rural or Urban
1	0.26	Background	Home Province
2	0.26	Pre-College	English First Additional
2	0.26	Pre-College	Geography
2	0.26	Background	Registration Start
2	0.26	Background	Registration End
2	0.26	Background	School Quintile
3	0.26	Pre-College	Computer Studies
3	0.21	Individual	Plan Code
3	0.21	Individual	Plan Desription
3	0.21	Individual	NBTQL
3	0.21	Individual	NBTAL

Table 7: Top 3 Principle Components and the Features which Produced those Components)

What is very clear from Table 7 is that the three main Tinto attributes form the base features of all three Principle Components (certain characteristics like Plan Code, School Quintile and Home Language are found to be characteristics of all components). This is a good indicator that the features of Tinto [1975] 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.

Final Feature Selection

To produce the final feature set I took the results of the Correlation Analysis (Table 5) and the Feature Importance (Table 6). When reviewing the tables together we can clearly see that there is an overlap in the most important features. Common features include Plan Code, Plan Description, School Quintile, Home Province, Mathematics, Science, NBT marks and English Home Language. Therefore, the final features are made up mainly using feature ranking produced from the Information Gain analysis with cross referencing to the Correlation analysis to determine if the correct attributes were indeed selected. The results of the PCA analysis (Table 7) were used to re-affirm that the final set of attributes really hold significant value in predicting the Risk Profile. All three feature selection techniques proved valuable and helped me build Table 8 below which is the final feature set used for modelling.

Tinto Category	**Feature**
Individual	PlanCode
Individual	PlanDescription
Individual	YearStarted
Individual	RegistrationStart
Individual	Majors
Individual	AgeatDropOutORGrad
Individual	NBTQL
Individual	NBTMA
Individual	NBTAL
Background	RegistrationEnd
Background	AgeAtFirstYear
Background	Homeprovince
Background	SchoolQuintile
Pre-College	MathematicsMatricMajor
Pre-College	LifeOrientation
Pre-College	PhysicsChem
Pre-College	AdditionalLanguage
Pre-College	Geography
Pre-College	LifeSciences
Pre-College	EnglishFirstAdditional

Table 8: Final Feature Set used for Modelling

Model Results and Evaluation

In this section I 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%.

Confusion Matrices

The 6 diagrams below represent the 6 different confusion matrices. Each confusion matrix represents the ratio of true positives, false positives, true negatives and false negatives for each of "Risk Profiles".

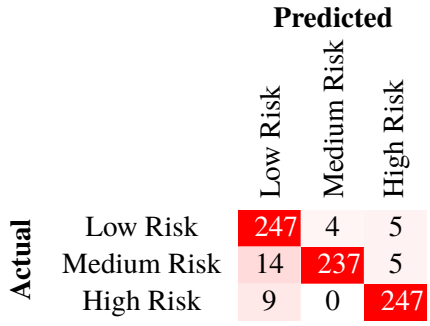

Figure 4: Random Forest - 95% Accuracy, Correctly identified 731/762 labels

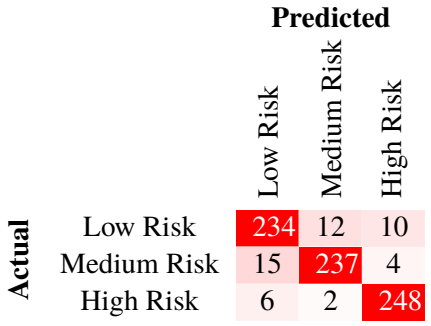

Figure 5: K-Star - 93% Accuracy, Correctly identified 719/762 labels

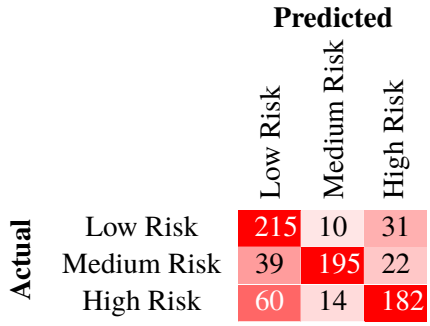

Figure 6: Naive Bayes - 78% Accuracy, Correctly identified 592/762 labels

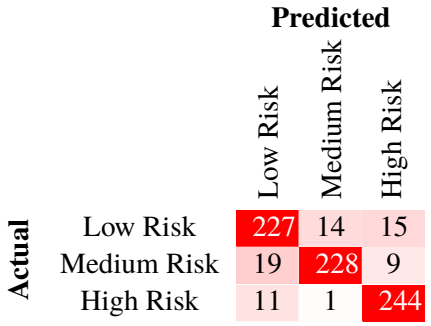

Figure 7: Multi-Layer Perceptron - 91% Accuracy, Correctly identified 699/762 labels

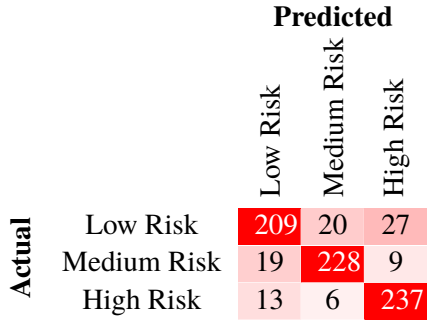

Figure 8: J-48 Algorithm - 88% Accuracy, Correctly identified 674/762 labels

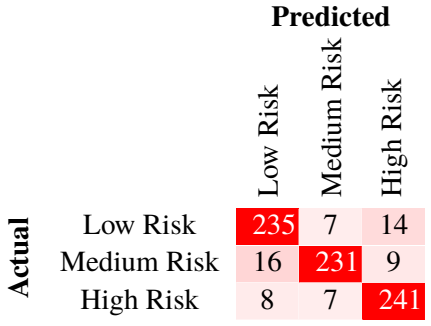

Figure 9: Logistic regression - 92% Accuracy, Correctly identified 707/762 labels

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%. 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 pre-emptive 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 Tinto [1975] - Background, Individual and Pre-College attributes can provide accurate indications of an individuals risk within a particular academic program.

Implications, Conclusion and Future Work

Given the current situation of Higher Education in South Africa - this includes 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 Background, Individual and Pre-College attributes of a student have proven to be successful in student attrition models by Tinto [1975], in that paper and others in literature. The Tinto [1975] model provided us with a solid conceptual framework to base our experiments of and true to the model; student risk profiles can be determined using the three features mentioned above.

This study looked at classifying a student into three different risk profiles when given a set of attributes of a student and a course plan code. The aim was to take these attributes and gauge the risk that a student would experience for a particular plan code. The categories included Low, Medium and High Risk students.

The dataset was obtained from the Academic Information Systems Unit (AISU). Permission from the AISU and Ethics departments (HREC Non-Medical) to use the data in this study was obtained (Certificate number: H19/08/28). The dataset used to produce the results contained 3 classes of size 256 samples per class. The complete dataset size was 768 records. The data was pre-processed synthetically using SMOTE to remove a class imbalance.

The feature selection process included getting the correlation, information gain (entropy) and a principal component analysis to the label value ("Risk Factor"). The three techniques used to evaluate the feature importance were then used to decide on a feature set that would be used for the modelling process

The modelling process consisted of 6 different modelling techniques. These techniques included a Random Forest, the K-Star Algorithm, Naive Bayes, a Multi-Layer Perceptron, the J-48 Decision Tree and lastly a Logistic Regression. The top three models were the Random Forest, K-Star and Logistic Regression with the Random Forest coming out on top with a 95% accuracy correctly identifying 731 out of 768 records. The sampling technique used for training and testing was 10 fold cross validation.

The Random Forest correctly classified 95% of the data whilst the other two were lower with averages of 93% and 92%. 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.

Bibliography

Adams, R. and Radix, C. (2018). Predicting student performance in a caribbean engineering undergraduate programme. 41:13–22.

Ajoodha, R. (2018). Representation, inference and learning. *IndabaX Conference*.

Ajoodha, R. (2019). Predicting learner attrition for the sciences using background, individual attributes and schooling at a sout african higher education reasearch intensive institution. Personal Communication.

Ajoodha, R. and Jadhav, A. (2019). Identifying at-risk undergraduate students using biographical and enrolment observations for mathematical science degrees at a south african university.

Ali, M., Qaseem, D. M., Rajamani, L., and Govardhan, D. (2013). Extracting useful rules through improved decision tree induction using information entropy. *International Journal of Information Sciences and Technology (IJIST)*, 3:27–41.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Campbell, P. and P. McCabe, G. (1984). Predicting the success of freshmen in a computer science major. *Commun. ACM*, 27:1108–1113.

Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W. (2002). Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357.

Chetty, Rajendra &Pather, S. (2015). *Telling stories differently. Engaging 21st century students through digital story telling*. Number Chapter: 1. SUN MeDIA Stellenbosch,.

Cleary, J. G. and Trigg, L. E. (1995). K*: An instance-based learner using and entropic distance measure. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, pages 108–114, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ezekiel, M. (1930). Methods of correlation analysis.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Ibrahim, Z. and Rusli, D. (2007). Predicting students' academic performance: Comparing artificial neural network, decision tree and linear regression. *21st Annual SAS Malaysia Forum*.

Jennifer M. Case, Delia Marshall, S. M. and Mogashana, D. (2018). *Going to University, The Influence of Higher Education on the Lives of Young South Africans*, volume 3. African Minds, 4 Eccleston Place, Somerset West 7130, Cape Town, South Africa.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Marock, C. (2008). Grappling with youth employability in south africa.

Martínez López, Y., Madera, J., and Varona, I. (2016). Study of the performance of the k* algorithm in international databases. *Revista Politécnica*, 12:2256–535.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math.

Osmanbegović, E. and Suljic, M. (2012). Data mining approach for predicting student performance. *Journal of Economics and Business Economic Review*, 10:3–12.

Peng, J., Lee, K., and Ingersoll, G. (2002). An introduction to logistic regression analysis and reporting.

Journal of Educational Research - J EDUC RES, 96:3–14.

Ramaswami, M. and Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining.

Romero, C., Ventura, S., Espejo, P., and Martínez, C. (2012). Data mining algorithms to classify students.

Smit, S. (2018). Sa unemployment on the rise - stats sa.

Thai-Nghe, N., Janecek, P., and Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. pages T2G–7.

Tinto, V. (1975). A synthetic model based on recent literature. *Review of Educational Research*, 45(1):89–125.