Forecasting Learner Attrition for Student Success at a South African University

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

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

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

ABSTRACT

In this paper we attempt to deduce student attrition at a South African higher-education institution with the aim of identifying students who are likely to be in need of academic support so that a focus could be provided on improving their academic performance. The significance of this paper is on using computer science and information technology to address learner attrition (an African reality) and thereby impact the low university throughput and retention rates positively.

We trained several machine learning classification models to deduce the student into four risk classes using only Grade 12 marks and background characteristics of the learner. We provide the following contributions: (a) the first known published trained classifier able to calculate the distribution over a students' risk profile for a South African university focused on the conceptual framework; (b) a ranking of employed features according to their entropy to correctly classify the class variable; (c) a comparison of trained classifiers able to calculate the probability of a students' risk profile for a South African higher-education research-intensive university; and (d) an interactive program which is able to calculate the posterior probability over the student's risk profile so that support can be provided to them.

The random forest classification model achieves the best performance with a 82% accuracy over these four risk profiles. We argue for introducing predictive tools to enhance student success and student support initiatives in Higher-Education Institutions. This work will benefit academic developers and staff who provide support to students who are at academic risk of completing their undergraduate Science programmes.

CCS CONCEPTS

• **Applied computing** → *Computer-assisted instruction*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Biographical characteristics, Background features, Individual attributes, Pre-college or schooling, Identifying at-risk biographical profiles, South Africa, Bayesian analysis, Model-based machine learning

ACM Reference Format:
Ritesh Ajoodha, Ashwini Jadhav, and Shalini Dukhan. 2020. 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 (SAICSIT '20), September 14–16, 2020, Cape Town, South Africa.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3410886.3410973

1 INTRODUCTION

Low attrition rates have prompted universities globally to increase efforts toward improving degree completion rates. One way of achieving this is through the use of data analytics to identify students that are struggling academically. Being able to predict student success is a necessary step in identifying academically challenged students so that an emphasis can be placed on support programmes and interventions aimed at enriching the student academic experience and improving their chances for academic success [1].

Additionally, the ability to predict the students who are at academic risk could prove useful in terms of resource allocation, sustained teaching efforts, and student support interventions. Universities usually invest significantly in student support programmes [42] and this cost is even more inflated for Science-related courses where provisions need to be made for laboratories, equipment and materials [15]. Therefore, insight into which classification models are most powerful in its prediction could help universities in their identification of undergraduates who would best benefit from support programmes.

Furthermore, within countries such as South Africa, where there is a need to address the historic imbalance caused by unfair constitutional approaches and practices of the past, it is necessary to provide wider access to higher-education for an improved quality of life and skills development [24]. However, when providing access to degree programmes, it is also crucial to establish a mechanism to enable the previously disadvantaged populations to gain epistemological access [26]. Being able to identify students early in the year who are most likely to benefit from support programmes has the potential to enrich the student's learning experience, and improve pass-rates and through-put at the university. An examination into

SAICSIT '20, September 14–16, 2020, Cape Town, South Africa

ACM ISBN 978-1-4503-8847-4/20/09…$15.00
https://doi.org/10.1145/3410886.3410973

the predictive power of classification models can assist in the early identification of students who are at academic risk.

Machine learning algorithms can be used to compare how different classification models could facilitate predictive power. The classification models enable the interpretation of factor interactions [9]. Comparisons across a range of models can elucidate which model has the highest degree of predictive power. Few studies have used a classification model to interpret the extent to which specific combinations of biographical and Grade 12 academic variables could best predict academic success at an undergraduate level in South Africa, nor has there been any known published work on which classification model could best predict specifically the academic success of Science undergraduates. This study fills these knowledge gaps, and would be useful to student support programmes.

This study focuses on a comparison of which combination of the following variables best predict academic success of undergraduates: the degree aspiration of the student (determined by the field in which the student has registered), year started, high school type, Grade 12 marks, performance over university benchmark tests, home province from which the student came, and country of origin. While some of these factors are associated with individual attributes and family background, others are related to the schooling background of the student.

[40] describes these three attributes (i.e. family background, individual attributes, and pre-college schooling) as factors related to attrition. In turn, according to Tinto's model, these three attributes influence the student's goal commitment and institutional commitment. Goal commitment and institutional commitment can be expanded on in terms of Academic System and Social System [40] (Figure 1), and the focus of this study lies squarely within the former system. As [40] explains, when examining drop-out within the field of education it is necessary to include individual characteristics such as school background, academic ability, as well as biographical characteristics. It is also necessary to take into account motivational attributes such as career aspirations.

While models in the past have been used to describe and provide explanations for drop-out, this paper advances existent literature by probing the use of classification models to predict the level of academic success in Science within undergraduate years in higher-education. The idea in this paper is to provide a support mechanism which academic developers and support programmes could use to identify, monitor and offer assistance to at-risk student cohorts in a focused and decisive way.

We define several features associated with Background or Family, Individual attributes, and Pre-college or Schooling which can be used to classify the student into four Risk Profiles: no risk, low risk, medium risk, and high risk. We trained several machine learning classification models such as decision trees; instance-based classifiers; naïve Bayes models; support vector machines; random forests; and linear logistic regression models to deduce the student into these four risk profiles. Confusion matrices were used to gauge model performance and factor analysis was performed to rate each feature's information gain to predicting the class value.

The results indicate that the student's chosen career path, Grade 12 Mathematics mark, the student's age at first year, and information about the school the student came from contributed the most towards correctly classifying student into a risk profiles. The best reported accuracy was the random forest algorithm which achieved 82% accuracy over the four classes. We note that in the random forest classification model, the miss-classifications occurred between the classes: no risk and low risk; low risk and medium risk; and medium risk and highest risk as reported in the reported confusion matrix. A web application has also been prepared which uses the random forest classification model to categorise a student's profile into these four risk classes using various school-leaving results and background characteristics.

We will provide the following contributions in this paper: (a) the first trained classifier able to calculate the probability of a student's Risk Profile for a South African Research-Intensive Higher-Education Institution focused around the conceptual framework of [40]; (b) a ranking of the student background, individual characteristics, and pre-college and schooling observations according to their information gain (entropy) to correctly classify the four classes; (c) a comparison of trained classifiers able to calculate the probability of a students' risk profile for a South African higher-education research-intensive university; (d) an interactive program which is able to calculate the posterior probability over these risk profiles given student background, individual characteristics, and high-schooling observations.

This document is structured as follows. The section 2 highlights the state-of-the-art contributions in the domain of predicting at-risk student profiles and a selected conceptual framework for student attrition; section 3 highlights our data, feature selection, and choice of classification models; section 4 outlines our major findings; and section 5 summarises this paper, outlines our contributions, and puts forward recommendations of future work.

2 RELATED WORK

The increasing access to the South African Higher-Education system has resulted in the admission of a large cohort of students from disadvantaged backgrounds [13]. While this is a positive trend, an investigative study on a cohort of university students in South Africa reported that the greatest attrition rate occurred at the end of the first year of study (29% of first year students). Moreover, only 30% of the total first-time student cohort had graduated after a five-year period [33]. This situation represents a crisis for higher-education systems and is a major problem in a country with limited state resources, which ultimately impacts the skill-base of the country [6]. The Student Pathways study by the Human Sciences Research Council also found that on average only 15% of students finish their degrees in the allotted time.

2.1 Conceptual framework

We adopt the conceptual framework of Tinto [40] as a rationale to predict student attrition using biographical and enrollment observations. [40] lists the following three input factors which contribute to student attrition as outlined in Figure 1: (a) Background or Family, (b) Individual Attributes, and (c) Pre-college or Schooling. These factors interrelate and influence the student's objective to complete their degree (goal commitment) or attitude towards university activities (institutional commitment). In the academic system, creating values and dispositions towards goal commitment translates to improved academic performance and intellectual development.

This leads to a decrease in the probability of dropping out [40]. The input factors of the conceptual framework put forward by [40], that is (a) Background or Family, (b) Individual Attributes, and (c) Pre-college or Schooling, indicates students family background, academic potential, and socio-economic status.

The input observation involved in the Tinto framework [40] deal with biographical and enrollment observations. These observations are seen to influence student attrition [1].

2.2 Challenges with transition into tertiary education

First-year students struggle to transit into the tertiary education system and have difficulty in adapting to the university environment as they find themselves deprived of the indispensable soft skills and much needed cultural capital for the pursuit of their studies [14]. An added challenge is often the inadequate level of education given at especially disadvantaged schools with majority of the students subsequently falling under the category of being "underprepared" [13]. Other determinants of academic performance include but are not limited to the students choice of study, when the student enters the system, the performance over benchmark tests, self-motivation, socio-economic status, age of student, learning preferences [8], class attendance [30], students' effort, previous schooling [34] and entry qualifications at universities.

The combination of factors influencing academic performance, however, varies from one academic environment to another, between institutions, between course of study, from one set of students to the next, and indeed from one cultural setting to another. Since not all factors are relevant for a particular context, it is imperative that specific studies be carried out to identify the context-specific determinants for early interventions to be effected in a timely manner.

This study was, therefore, designed to identify and analyse few determinants or markers that help identify at-risk students at a particular university in South Africa using biographical and enrolment observations for Science programmes. Statistically, the success of a student in a degree or course is correlated with their performance in previous assessments and their biographical associations (i.e. school quintile, home province, and school type). In this paper we explore identifying 'at-risk' students using biographical and enrolment observations. More specifically, this research will focus on diagnosing the biographical profiles of students whose academic trajectory will potentially lead to failing to complete the requirements of their degree. It is believed that if students at-risk of completing their studies are identified early in the year, then support measures and monitoring systems can be instituted promptly and timeously to improve the students chances of academic success.

The student attrition has significant consequences for student, academic, administrative staff, parents and communities. The electronic storage of student data and ability to slice and dice based on various parameters have made the educators understand various factors affecting the student success through data mining techniques. These predictors are now being assembled for forecasting the success of not only students but also education programmes. Predictive models use statistics to predict outcomes based on historical data.

[32] and [28] summarised the influencing predictors as admission student centric attributes (age, ethnicity, prior academic performance, gender, time management, information / computer literacy, reading and writing capability); after admission academic centric attributes (academic integration, clarity of programs, interpersonal relationship, study habits, accessibility to services, absenteeism, social commitment etc.) and external environment attributes (financial status, family responsibilities, external support, life crisis etc.,) that influence the students' decisions.

In the last ≈ 25 years the analysis of student data has transformed from mining of survey data [31], use of principal component analysis by [7] to use of machine learning to predict student dropouts by [38]. In one study of 450 students who enrolled in 71150 information management courses over three years (2006-2009), [22] showed that socio-demographic predictors like ethnicity, course programme and course were significant predictors of students being successful or not. He also showed that the level of risk estimated based on enrolment data is not enough to predict the student's fate [43] and [36] found for courses in mathematics and computing that the biggest factors for student attrition were the marks in the first assignment, course level followed by course rating, course work, gender, age and socio-economic status of the students.

[21] compared various supervised machine learning algorithms like decision trees, artificial neural networks, naïve Bayes classifier, instance-based learning, logistic regression and support vector machines to predict student's success. These authors showed that when only demographic predictors are used on the neural networks, the accuracy was 58.84 %, but there is an increased accuracy to 64.47% when using support vector machines. However, the authors additionally indicated that when other predictors were used along with demographic factors, the naïve Bayes classifier accurately predicted performance of the student. This paper informed the model selection of this research.

[41] showed that decision trees, neural networks and linear discriminant analysis failed to predict academic success using demographic data and academic history of students in Belgian-French speaking universities. They also conclude that previous education, prior mathematics course work, and age were significant predictors of student success whereas gender, parents education, marital status played no role in student success. Our paper adds to extant literature by showing how demographic and academic data can improve the predictive power of classification models, and the usefulness of the classifier in identifying at-risk students who could benefit from more focused academic support early on in the academic year. In the next section we will present the methodology explored by this paper.

3 METHODOLOGY

In this paper we attempt to use Grade 12 marks and some biographical characteristics to predict the distribution over several risk profiles. We use the following risk profiles in this study: *"No Risk"*, where the student completes their degree in minimum time (3 years); *"Low risk"*, where the student completes their degree in more than the minimum time; *"Medium risk"*, where the student fails their degree before the minimum time of completion; and *"High risk"*, where the student fails their degree and exceeds the

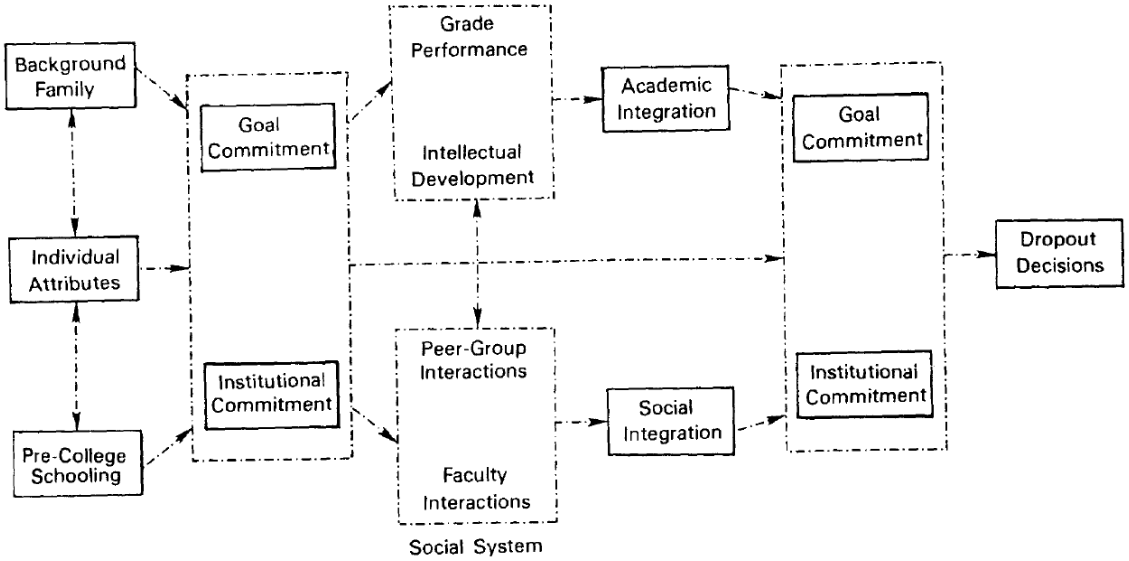


Figure 1: The graphical user interface adapted for use for the at-risk program, Tinto (1975) (sic).

minimum time. We differentiate between Medium risk and High risk profiles since the High risk profile uses more of the student's and university resources towards a failed outcome.

We trained several machine learning classification models from different archetypes of machine learning to deduce the student into one of these four Risk Profiles. Confusion matrices will be used to gauge model performance and factor analysis will be performed to rate each features contribution to predicting the class label.

This section is structured as follows: We firstly describe the data collection and pre-processing steps taken to prepare the data for this research objective; secondly we outline the features used to predict the class variables as well as a mechanism to gauge the contribution of each feature used; thirdly we provide brief descriptions of the machine learning classifiers used to perform the classification and evaluation metrics; and finally, we provide information concerning the ethics clearance certificate details.

3.1 Data Collection and Pre-processing

The data used in this study consisted of biographical and enrolment observations of students from the Faculty of Science degrees at a South African university. These degrees include streamlines of specialities from four major science streamlines: Earth Sciences, Biological and Life Sciences, Physical Sciences, and Mathematical Sciences. The enrolment and biographical observations were collected for all students registered anytime between the years 2008 to 2018.

3.2 Features used and Information Gain

As explained earlier, we adopt the conceptual framework of [40] to develop a methodology to predict student attrition using biographical and enrolment observations. Firstly, for (a) Background and Family, we used the following features: the country and province/state from which the student originated; the quintile associated with the school from which the student came; whether the school is an urban or rural school (school type); and the age of the student at first year. Secondly, for (b) Individual Attributes, we attempted to find a measure of the students' proficiency to understand academic literacy, quantitative literacy, and mathematical literacy to meet the demands of university-level work. We also used the national benchmark tests NBTAL, NBTQL, and NBTMA respectively to inform us about the individual attributes of the learner. We also considered the students' intended plan description as a fair indication of what the student aspires towards for their professional career. Finally, for (c) Pre-college or Schooling, we considered the students' school-leaving results for the following subjects: Life Orientation; Core Mathematics; Mathematics Literacy; Additional Mathematics; English Home Language; Physical Science; English First Additional; and Computer Studies. No university semester marks were used in this study.

There are many mentioned attributes which contribute to the successful prediction of a student's Risk Profile. However, not all of these observations are equally meaningful for this particular task. The problem of feature selection is broken up into two components: (a) declaring a mechanism to perform feature evaluation with respect to the class variable (Risk Profile), and (b) using a feature evaluator to navigate combinations of features to derive the information loss of using variables subsets of the feature list. We will use the Information Gain Ranking (IGR) algorithm to perform feature analysis. The IGR algorithm calculates the entropy (information gain) for each feature with respect to the class variable. The entropy, $0 \geq e \leq 1$, where 0 indicates no information gain and 1 indicates maximum information gain. In the next section we will describe the machine learning classifiers used in this paper.

3.3 Classification and Evaluation

We use the following six off-the-shelf classification models to predict the Risk Profile of a student: Decision trees, K-Star, Naïve Bayes, Support Vector Machines (SVMs), Random Forests, and Linear Logistic Regression Models.

Decision trees. The decision tree algorithm we selected for this task was the $C4.5$ classification model. The $C4.5$ algorithm uses information entropy to build a decision tree based on the ID3 algorithm [18]. The $C4.5$ algorithm recursively selects a feature with the greatest information gain to split the training sample. This intuitively allows the most important feature, with respect to the class variable, to make the 'decisions' from the root down the tree. The $C4.5$ classification procedure implemented in this paper follows the original algorithm by [29].

K^*. The K* instance-based classifier uses an entropy-based distance function to classify test instances using the training instance most similar to them. The K* implementation used in this paper closely followed the implementation by [12]. Using an entropy-based distance function allows consistency in the classification of real-valued and symbolic features found in our experiments.

The Naïve Bayes Model. Perhaps the simplest example of a Bayesian model is the naïve Bayes model (NBM) which has been traditionally and successfully used by many expert systems [3, 20]. The NBM pre-defines a finite set of mutually exclusive classes. Each instance can fall into one of these classes, this is represented as a latent class variable. The model also poses some observed set of features $X_1, \ldots, X_n$. The assumption is that all of the features are conditionally independent given the class label of each instance [5]. That is $\forall \mathbf{i}(\mathbf{X_i} \perp\!\!\!\perp \mathbf{X_{i'}} \mid \mathbf{C})$, where $X_{i'} = \{X_1, \ldots, X_n\} - \{X_i\}$. The joint distribution of the NBM factorises compactly as a prior probability of an instance belonging to a class, $P(C)$, and a set of conditional probability distributions (CPDs) which indicate the probability of a feature given the class. We can state this distribution more formally as $P(C, X_1, \ldots, X_n) = P(C) \prod_{i=1}^{n} P(X_i|C)$. [2] provides a comprehensive introduction to Bayesian methods.

The NBM remains a simple, yet highly effective, compact, and high-dimensional probability distribution that is often used for classification problems [4]. The implementation of the NBM follow that of [19].

Support Vector Machines. The Support Vector Machine (SVM) classification model incorporates the training data into a non-probabilistic binary linear classifier which separates the classes of the training data by a multi-dimensional hyper-plane. Test instances are then mapped on the same space and predicted based on which side of the hyper-plane they fall on. Using a kernal trick and the one-verses-all class partitioning, SVMs can be scaled for nonlinear and high-dimensional classification. The SVM implementation used in this paper follows the implementation by [11, 16].

Random Forests. Random decision forests are an ensemble classification learning method that uses the training data to build several decision trees based on the mode of the class variable. This technique of using several decision trees prevents over-fitting compared to a single decision tree. The implementation used in this paper is based on [10].

Linear Logistic Regression Models. The Linear Logistic Regression classification model uses additive logistic regression as mentioned in [17] with added simple regression functions as base learners [3]. The implementation used in this paper follows [23, 37].

All six of these classification models will be evaluated using a confusion matrix [39] and the associated classification accuracy will be provided alongside each model. A 10-fold cross validation scheme will be used [44].

3.4 Ethics Clearance

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

4 RESULTS

In this section we present the results of performing machine learning classification to predict student risk profiles using background and school leaving results. This section is structured as follows: first we present the results of the feature analysis; and thereafter we present the classification results.

4.1 Feature Information Gain

There were 20 features used in this paper to predict the class variable. Using IGR we could deduce the contribution of each feature to classify the features as a value from the class variable. Table 1 illustrates a ranking of the contribution of each feature to classify the Risk Profile.

The first column indicates the rank of the feature from most contributing feature (rank 1) to least contributing feature (rank 20); column 2 indicates the entropy value associates with each feature where $0 \leq e \geq 1$ (0 no information to 1 maximum information); and the third column indicates the feature name/description. The feature are colour coded relating them to the [40] conceptual framework (brown indicating background or family, blue indicating individual attributes, and black indicating pre-collage or schooling).

Although the ranking of the feature set using entropy provides a useful framework for feature elimination, the entropy value also indicates the contribution of each feature relative to the other features. Figure 2 illustrates a plot of the entropy values as organised in Table 1. As the function in Figure 2 monotonically increases, the loss in entropy between each subsequent point decreases logarithmically. The seven most contributing features from Table 1 are highlighted.

4.2 Classification

In this section we will present the result of the classification algorithms. The following six classification procedures where employed in this paper: Decision trees, K*, Naïve Bayes, SVMs, Random forests, and Linear Logistic Regression Models. Figure 3 indicates the result of each of these classifiers to predict the class variable.

Figure 3a illustrates the confusion matrix for the $C4.5$ classification model which achieves 79% accuracy using 10-fold cross validation. With the exception of the K* and naïve Bayes, compared

Rank	Entropy	Feature Name
1	1.21960228	PlanCode
2	1.15086266	PlanDescription
3	0.59886383	Streamline
4	0.29582771	Year Started
5	0.20836689	AgeatFirstYear
6	0.18695721	SchoolQuintile
7	0.14234042	MathematicsMatricMajor
8	0.12166049	Homeprovince
9	0.06417526	isRuralorUrban
10	0.0568866	LifeOrientation
11	0.04978826	PhysicsChem
12	0.02780914	EnglishFirstLang
13	0.01253064	Homecountry
14	0.00550434	AdditionalMathematics
15	0.00000902	MathematicsMatricLit
16	< 0.00001	NBTAL
17	< 0.00001	NBTMA
18	< 0.00001	NBTQL
19	< 0.00001	ComputerStudies
20	< 0.00001	EnglishFirstAdditional

Table 1: A ranking of the information gain (entropy) for a set of features to predict the students Risk Profile (class variable). The top seven features are highlighted indicating entropy greater than 0.1.

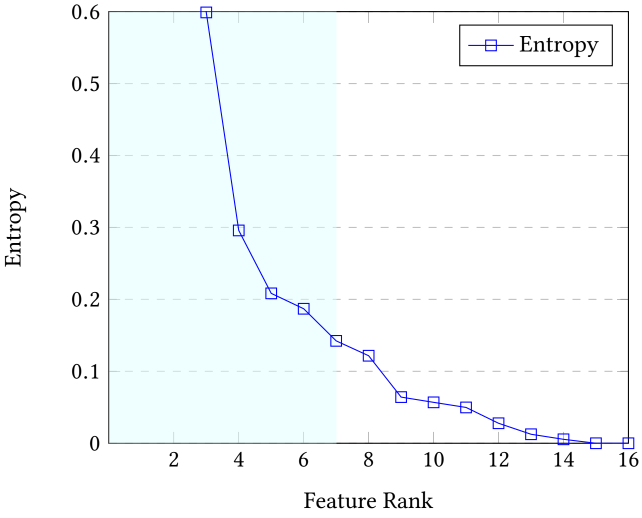


Figure 2: A graphical illustration of the information gain (entropy) for a set of features to predict the students Risk Profile (class variable). The x-axis indicates the feature rank and the y-axis indicates the information gain for using that feature.

to the other four classification models employed in this paper the $C4.5$ took the least time to build.

Figure 3b illustrates the confusion matrix for the K* classification model which achieves 75% accuracy using 10-fold cross validation. From all six classification models employed in this paper K* took the least time to build.

Figure 3c illustrates the confusion matrix for the naïve Bayes classification model which achieves 80% accuracy using 10-fold cross validation. With the exception of the K* classification model, from the other five models employed in this paper naïve Bayes took the least time to build.

Figure 3d illustrates the confusion matrix for the SVM classification model which achieves 52% accuracy using 10-fold cross validation, the worst classification accuracy achieved in this paper. Furthermore, from all six classification models employed in this paper the SVM classification model took the longest time to build.

Figure 3e illustrates the confusion matrix for the random forest classification model which achieves 82% accuracy using 10-fold cross validation, the highest classification accuracy achieved in this paper. With the exception of the K*, Naïve Bayes, and $C4.5$, compared to the other three classification models employed in this paper the random forest model took the least time to build.

Figure 3f illustrates the confusion matrix for the linear logistic regression classification model which achieves 78% accuracy using 10-fold cross validation. With the exception of the SVM classification model, the Linear Logistic Regression Model took the longest time to build.

In the next section, we interpret these results and provide a web application using the random forest classification model that can predict the at-Risk Profile of the student based on demographic and academic factors.

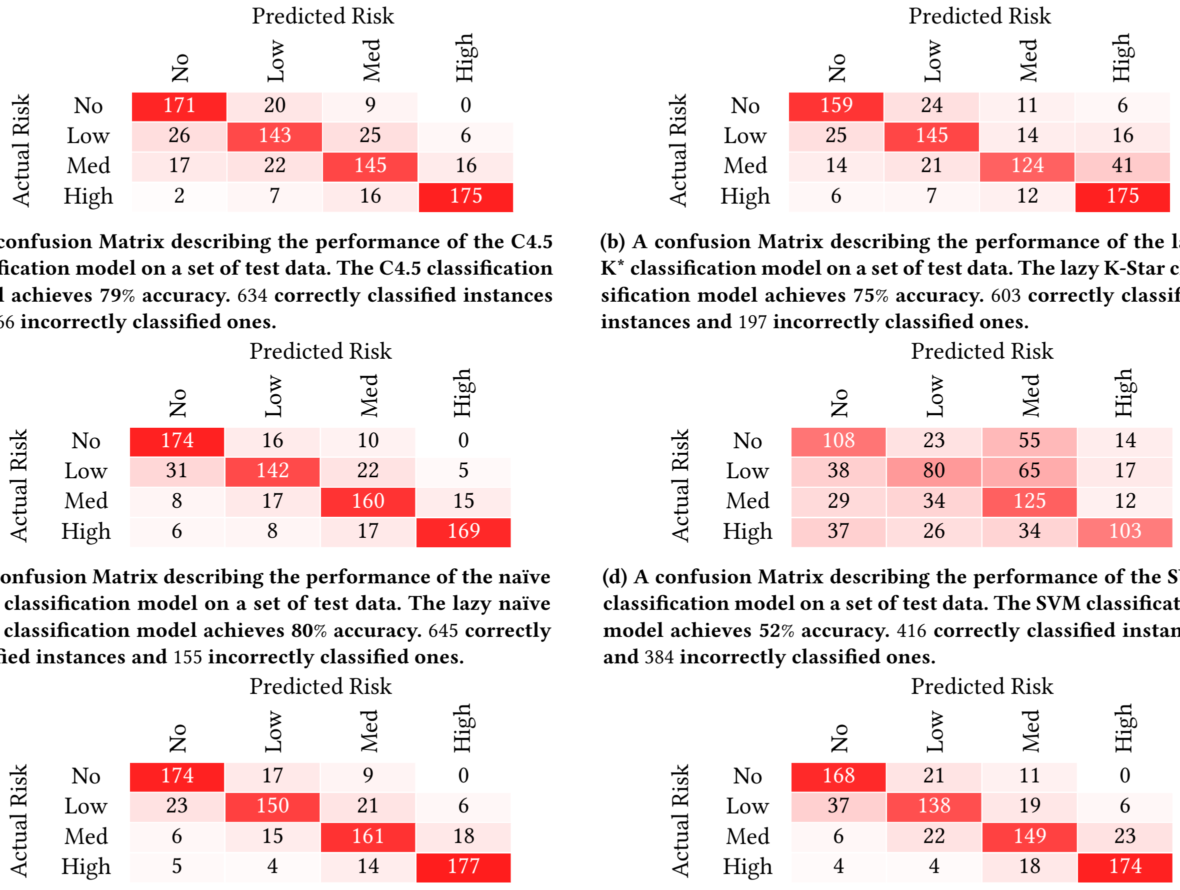
5 DISCUSSION AND CONCLUSION

Although the combined feature set achieves 82% accuracy over the four Risk Profiles using the random forests, not all the listed features in Table 1 provide an equal contribution towards correctly classifying the class variable. While the graph in Figure 2 monotonically decreases, the contribution of each subsequent feature is similar after the seventh rank, which means that we lose an increasingly smaller entropy with every employed feature. An example of the practical implication of this is that trading off Additional Mathematics for Computer Studies will result in less lost entropy ($e \approx 0.001$) rather than trading off Year Started for School Quintile ($e \approx 0.2$).

We note that the random forest classification model outperformed the other five models and provided the fastest build time with the exception of the K* and naïve Bayes models. In terms of interpreting the recorded incorrectly classified instance, we observed the severity of the misclassifications indicated in Figure 3. For example the misclassification of instances by the SVM as 27% of No Risk instances being incorrectly classified as Medium Risk is far more severe than the misclassification of instances by the random forest of 5% of No Risk instances being incorrectly classified as Medium risk, given the definition of the class labels and sensitivity of instances being incorrectly classified.

5.1 Contribution

We provide an automated system to predict the Risk Profile of a student based the input features from the conceptual framework of [40]: background or family; individual attributes; and pre-college



Actual Risk \ Predicted Risk	No	Low	Med	High
No	171	20	9	0
Low	26	143	25	6
Med	17	22	145	16
High	2	7	16	175

(a) A confusion Matrix describing the performance of the C4.5 classification model on a set of test data. The C4.5 classification model achieves 79% accuracy. 634 correctly classified instances and 166 incorrectly classified ones.

Actual Risk \ Predicted Risk	No	Low	Med	High
No	159	24	11	6
Low	25	145	14	16
Med	14	21	124	41
High	6	7	12	175

(b) A confusion Matrix describing the performance of the lazy K* classification model on a set of test data. The lazy K-Star classification model achieves 75% accuracy. 603 correctly classified instances and 197 incorrectly classified ones.

Actual Risk \ Predicted Risk	No	Low	Med	High
No	174	16	10	0
Low	31	142	22	5
Med	8	17	160	15
High	6	8	17	169

(c) A confusion Matrix describing the performance of the naïve Bayes classification model on a set of test data. The lazy naïve Bayes classification model achieves 80% accuracy. 645 correctly classified instances and 155 incorrectly classified ones.

Actual Risk \ Predicted Risk	No	Low	Med	High
No	108	23	55	14
Low	38	80	65	17
Med	29	34	125	12
High	37	26	34	103

(d) A confusion Matrix describing the performance of the SVM classification model on a set of test data. The SVM classification model achieves 52% accuracy. 416 correctly classified instances and 384 incorrectly classified ones.

Actual Risk \ Predicted Risk	No	Low	Med	High
No	174	17	9	0
Low	23	150	21	6
Med	6	15	161	18
High	5	4	14	177

(e) A confusion Matrix describing the performance of the random forests classification model on a set of test data. The random forests classification model achieves 83% accuracy. 662 correctly classified instances and 138 incorrectly classified ones.

Actual Risk \ Predicted Risk	No	Low	Med	High
No	168	21	11	0
Low	37	138	19	6
Med	6	22	149	23
High	4	4	18	174

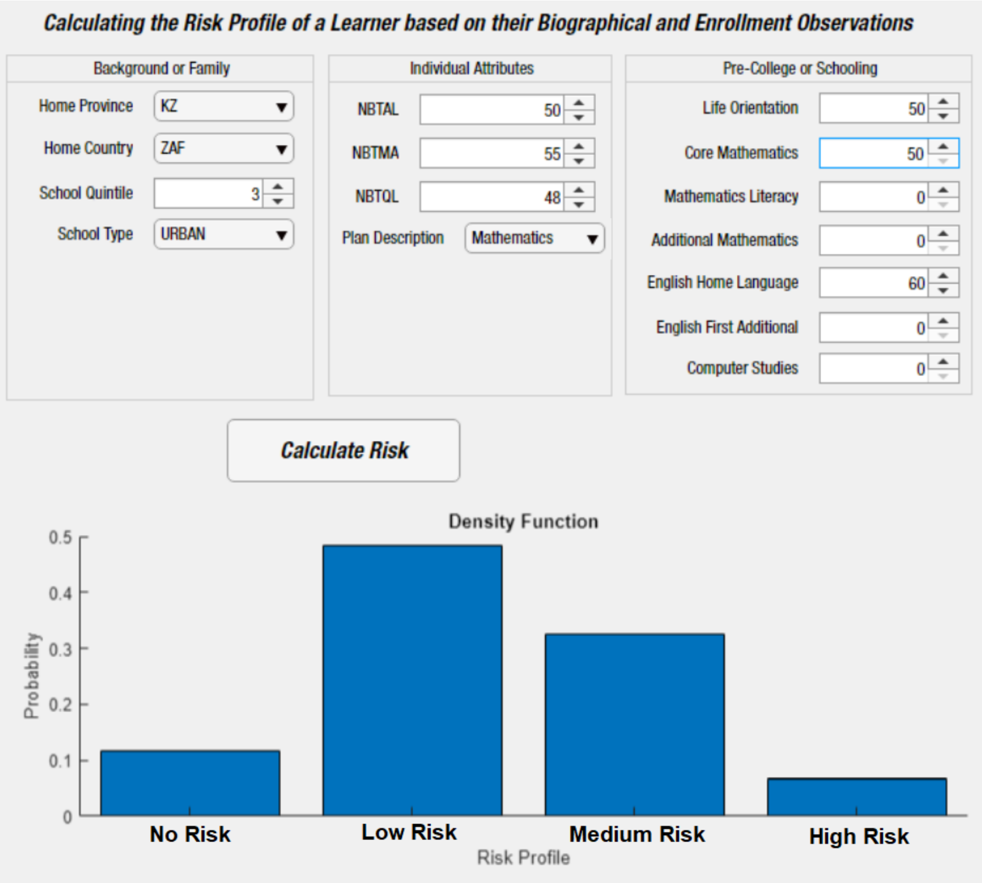
(f) A confusion Matrix describing the performance of the linear logistic regression classification model on a set of test data. The linear logistic regression classification model achieves 78% accuracy. 629 correctly classified instances and 171 incorrectly classified ones.

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

or schooling. The rationale for this aim is to provide a practical tool which academic developers and support programmes could use for early identification of students who are in need of academic assistance at university. The early identification and monitoring of students who are likely to be at academic risk can lead to increased pass-rates and through-puts at university, and provide the student with a more enriched learning experience in first year.

This research points to the importance of age of the learner at first year as a possible indication of maturity; the prominence of the high school attended by the learner (quintile and type); and the importance of performance in Core Mathematics in Grade 12 to predict the success of a learner.

As [24] explains, higher-education institutions need to provide students with the tools to effectively participate in this academic environment. However, in large class situations which is often the case in the undergraduate years, it is difficult to monitor the progress of individual students and provide the necessary support for all. [25] and [27] also evidence that there is a growing body of first year students in the developing context (i.e. countries such as South Africa) who are under-prepared for university. The application which we suggest in this paper can assist university academic developers and support programmes identify students who are mostly likely in need of academic intervention. This level of focused effort could enable universities to create a level platform which enables the academic success of previously disadvantaged students. Although research has indicated [35] that the efficacy of support programmes need investigation, there is also a need to understand how support programmes could more effectively select for students who could benefit from participation. Figure 4 provides an example of how the tool in this study was used to analyse data.



Calculating the Risk Profile of a Learner based on their Biographical and Enrollment Observations

Background or Family

- Home Province: KZ
- Home Country: ZAF
- School Quintile: 3
- School Type: URBAN

Individual Attributes

- NBTAL: 50
- NBTMA: 55
- NBTQL: 48
- Plan Description: Mathematics

Pre-College or Schooling

- Life Orientation: 50
- Core Mathematics: 50
- Mathematics Literacy: 0
- Additional Mathematics: 0
- English Home Language: 60
- English First Additional: 0
- Computer Studies: 0

Figure 4: The graphical user interface for the at-risk program.

5.2 Application

In Figure 4 features are categorised into the input features as set out by [40]. The example in the figure attempts to reveal the predicted posterior distribution over the four Risk Profiles (No risk, Low Risk, Medium Risk, and High Risk) using a student with the following observations: A student from KwaZulu-Natal in South Africa; from an urban school with a quintile of 3; with a 50%, 55%, and 48% score in the National Benchmark tests on Academic Literacy, Mathematical Literacy, and Quantitative Literacy respectively; achieving a 50%, in Life Orientation; 50% in Core Mathematics; and 60% in English Home Language. The output of the program is that the student is hypothetically 10% likely to be at No Risk (Completion of degree in 3 years); 50% likely of being at Low Risk (Completion in greater than 3 years); 35% likely to be at Medium Risk (Dropout before 3 years); and 5% likely to be at High Risk (dropout in greater than 3 years). The underlying model used to perform the prediction task in the app in Figure 4 is the random forest which gave the best classification accuracy from the models used in this paper. Information such as this could prove to be useful to academic developers and support programmes in terms of identifying which students are more likely to experience challenge in the academic environment at university, and thus focus their efforts on helping students who are in a particular risk category.

5.3 Recommendation and Implication of Research

Degree programmes in the Sciences at universities often focus on methods and approaches to improve academic achievement of students, this is directly tied in with producing graduates who are globally competitive when they enter into the workforce.

The application that has been proposed in this paper can assist in identifying students who are likely to benefit from support programmes aimed at improving academic achievement early in the academic year. The application is appropriate for different contexts since it accounts for biographical and academic variables, thus academic developers or programme co-ordinators will be able to use the application for early identification of students from a range of backgrounds who are considered in need of support.

The implication of this recommendation is that students will be supported early in their academic career, thus increasing their chance of academic success than 1) if they were to self-identify for support programmes, and 2) if support measures were put into place later in the year, i.e. only when test results are released.

5.4 Future Research

Future avenues of research can (a) explore the impact of the highly ranked features in Table 1 of student attrition; (b) model the student in the academic system of [40] through their grade performance and intellectual development; or (c) incorporate observations which reveal the Peer-Group Interactions and Faculty Interactions in the Social System of [40].

5.5 Limitations

There are several data limitations and biases that confine this study. Firstly, the latent factors which influence the quantity and diversity of the background observations are not thoroughly understood. There could be confounding factors not included in this study which may provide higher predictive accuracy. The results of this paper are based solely from the observations recorded in the provided data. Secondly, the predictive accuracy relies on the quality and amount of the recorded data and future studies are required to gauge the consistency and generality of the results derived from the South African Research-Intensive Higher-Education Institution used by this paper. Thirdly, the use of socio-cultural and socio-economic factors were not included in this paper, incorporating these factors may allow us to integrate the Social System [40] into the predictive model which could have provided a greater information gain from the data. This paper opens the possibilities to explore these avenues by providing a quantitative motivation and means to identify at-risk student profiles.

ACKNOWLEDGMENTS

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant numbers: 121835 and 121960).

REFERENCES

[1] Tasneem Abed, Ritesh Ajoodha, and Ashwini Jadhav. 2020. A Prediction Model to Improve Student Placement at a South African Higher Education Institution. In *2020 International SAUPEC/RobMech/PRASA Conference*. IEEE, 1–6.

[2] Ritesh Ajoodha. 2019. Influence modelling and learning between dynamic bayesian networks using score-based structure learning. *The University of the Witwatersrand, Johannesburg (wirespace)* (2019).

[3] Ritesh Ajoodha, Richard Klein, and Benjamin Rosman. 2015. Single-labelled music genre classification using content-based features. In *2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*. IEEE, 66–71.

[4] Ritesh Ajoodha and Benjamin Rosman. 2017. 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)*. IEEE, 122–127.

[5] Ritesh Ajoodha and Benjamin Rosman. 2018. Learning the influence structure between partially observed stochastic processes using IoT sensor data. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

[6] D Andrews and R Osman. 2015. Redress for academic success: possible'lessons' for university support programmes from a high school literacy and learning intervention: part 2. *South African Journal of Higher Education* 29, 1 (2015), 354–372.

[7] Francisco Araque, Concepción Roldán, and Alberto Salguero. 2009. Factors influencing university drop out rates. *Computers & Education* 53, 3 (2009), 563–574.

[8] Rasimah Aripin and Zurina Mahmood. 2008. *Students' learning styles and academic performance*.

[9] Cédric Beaulac and Jeffrey S Rosenthal. 2019. Predicting University Students' Academic Success and Major Using Random Forests. *Research in Higher Education* (2019), 1–17.

[10] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.

[11] Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM - A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/ The Weka classifier works with version 2.82 of LIBSVM.

[12] John G. Cleary and Leonard E. Trigg. 1995. K*: An Instance-based Learner Using an Entropic Distance Measure. In *12th International Conference on Machine Learning*. 108–114.

[13] Michael Cross and Claude Carpentier. 2009. 'New students' in South African higher education: institutional culture, student performance and the challenge of democratisation. *Perspectives in Education* 27, 1 (2009), 6–18.

[14] Shalini Dukhan, Ann Cameron, and Elisabeth A Brenner. 2012. The Influence of Differences in Social and Cultural Capital on Students' Expectations of Achievement, on their Performance, and on their Learning Practices in the First Year at University. *International Journal of Learning* 18, 7 (2012).

[15] Gary Leo Dunbar. 2019. Strategies to maximize the involvement of undergraduates in publishable research at an R2 University. *Frontiers in Psychology* 10 (2019), 214.

[16] Yasser EL-Manzalawy. 2005. WLSVM. http://www.cs.iastate.edu/~yasser/wlsvm/ You don't need to include the WLSVM package in the CLASSPATH.

[17] J. Friedman, T. Hastie, and R. Tibshirani. 1998. *Additive Logistic Regression: a Statistical View of Boosting*. Technical Report. Stanford University.

[18] Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, and Mohammed Erritali. 2014. A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications* 4, 2 (2014), 0–0.

[19] George H. John and Pat Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Mateo, 338–345.

[20] Dyna Marisa Khairina, Septya Maharani, Heliza Rahmania Hatta, et al. 2017. Decision Support System for Admission Selection and Positioning Human Resources by Using Naive Bayes Method. *Advanced Science Letters* 23, 3 (2017), 2495–2497.

[21] Sotiris Kotsiantis, Christos Pierrakeas, and Panagiotis Pintelas. 2004. PREDICTING STUDENTS'PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES. *Applied Artificial Intelligence* 18, 5 (2004), 411–426.

[22] Zlatko Kovacic. 2010. Early prediction of student success: Mining students' enrolment data. (2010).

[23] Niels Landwehr, Mark Hall, and Eibe Frank. 2005. Logistic Model Trees. 95, 1-2 (2005), 161–205.

[24] Henry D Mason. 2019. Evaluation of a study skills intervention programme: A mixed methods study. *Africa Education Review* 16, 1 (2019), 88–105.

[25] Monkie Moseki and Salome Schulze. 2010. Promoting self-regulated learning to improve achievement: A case study in higher education. *Africa Education Review* 7, 2 (2010), 356–375.

[26] Johan Muller. 2014. Every picture tells a story: Epistemological access and knowledge. *Education as Change* 18, 2 (2014), 255–269.

[27] C Nel, C Troskie-de Bruin, and E Bitzer. 2009. Students' transition from school to university: Possibilities for a pre-university intervention. *South African Journal of Higher Education* 23, 5 (2009), 974–991.

[28] Ji-Hye Park. 2007. Factors Related to Learner Dropout in Online Learning. *Online Submission* (2007).

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

[30] David Romer. 1993. Do students go to class? Should they? *Journal of Economic Perspectives* 7, 3 (1993), 167–174.

[31] Cristobal Romero and Sebastian Ventura. 2007. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications* 33, 1 (2007), 135–146.

[32] Alfred P Rovai. 2003. In search of higher persistence rates in distance education online programs. *The Internet and Higher Education* 6, 1 (2003), 1–16.

[33] Ian Scott, Nan Yeld, and Jane Hendry. 2007. A case for improving teaching and learning in South African higher education. *Higher education monitor* 6, 2 (2007), 1–8.

[34] John J Siegfried and Rendigs Fels. 1979. Research on teaching college economics: A survey. *Journal of economic literature* 17, 3 (1979), 923–969.

[35] TD Sikhwari and J Pillay. 2012. Investigating the effectiveness of a study skills training programme. *South African Journal of Higher Education* 26, 3 (2012), 606–622.

[36] Ormond Simpson. 2006. Predicting student success in open and distance learning. *Open Learning: The Journal of Open, Distance and e-Learning* 21, 2 (2006), 125–138.

[37] Marc Sumner, Eibe Frank, and Mark Hall. 2005. Speeding up Logistic Model Tree Induction. In *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Springer, 675–683.

[38] Mingjie Tan and Peiji Shao. 2015. Prediction of student dropout in e-learning program through the use of machine learning method. *International Journal of Emerging Technologies in Learning (iJET)* 10, 1 (2015), 11–17.

[39] Kai Ming Ting. 2017. Confusion matrix. *Encyclopedia of Machine Learning and Data Mining* (2017), 260–260.

[40] Vincent Tinto. 1975. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research* 45, 1 (1975), 89–125.

[41] J-P Vandamme, Nadine Meskens, and J-F Superby. 2007. Predicting academic performance by data mining methods. *Education Economics* 15, 4 (2007), 405–419.

[42] Lynette Vernon, Stuart J Watson, William Moore, and Sarah Seddon. 2019. University enabling programs while still at school: supporting the transition of low-SES students from high school to university. *The Australian Educational Researcher* 46, 3 (2019), 489–509.

[43] R Woodman. 2001. Investigation of factors that influence student retention and success rate on Open University courses in the East Anglia region. *Unpublished M. Sc. Dissertation), Sheffield Hallam University, UK* (2001).

[44] Ping Zhang. 1993. Model selection via multifold cross validation. *The Annals of Statistics* (1993), 299–313.