2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC) | 978-1-7281-9520-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/IMITEC50163.2020.9334103

A Conceptual Model to Identify Vulnerable Undergraduate Learners at Higher-Education Institutions

Noluthando Mngadi
School of Computer Science
and Applied Mathematics
University of the Witwatersrand,
Johannesburg, South Africa
noluthando.mngadi@gmail.com

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

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

***Abstract*—There is a growing concern around student attrition worldwide, including South African universities. More often than not, the reasons for students not completing their degree in the allocated time frame include academic reasons, socio-pschyo factors, and lack of effective transition from the secondary education system to the tertiary education systems. To overcome these challenges, the tertiary educational institutions endeavor to implement interventions geared toward academic success. One of the challenges, however, is identifying the vulnerable students in a timely manner. This study therefore aims to predict student performance by using a learner attrition model so that the vulnerable students are identified early on in the academic year and are provided support through effective interventions, thereby impacting student success positively.**

Predictive machine learning methods, such as support vector machines, decision trees, and logistic regression, were trained to deduce the students into four risk-profiles. A random forest outperformed other classifiers in predicting at-risk student profiles with an accuracy of 85%, kappa statistic of 0.7, and an AUC of 0.95. This research argues for a more complex view of predicting vulnerable learners by including the student's background, individual, and schooling attributes.

***Index Terms*—Attrition, At-risk, Machine learning, Classification.**

I. Introduction

Many students in higher-education institutions struggle to complete their undergraduate curriculum on time. This is due to various factors. To address this issue, many educational institutions provide numerous intervention programmes, either on an individual or group basis, to promote student success. In particular, first-year students are affected as they have not yet conformed to the culture of the academic system [1], [2]. Interventions can be in the form of support programs that are intended to be applied to assist students in coping with their academic experience. To best apply the intervention programmes, higher-education institutions need to identify vulnerable learners as early as possible to allow enough time for their integration back into the academic system.

In a report by Michael and Susan Dell Foundation in 2017, about 32% of students who are financially supported complete a 3-year degree after 5-years [3]. Identifying and remediating vulnerable learners is of great benefit to many stakeholders. These include the student, the lecturer, the university, and other sponsors of the learner. Many studies define vulnerable learners as those who drop-out or fail-out of a course or programme [4]. This study will adopt the definition of vulnerable learners as those whose interrelations of biographical, individual, and schooling characteristics have a higher probability of failing to meet the minimum requirements to obtain an undergraduate degree in record time (3-years) [5].

This paper explores the relationship between background, individual, and schooling characteristics on learner attrition, as per the learner attrition model proposed by [6]. This paper is motivated by the underlying themes of student support and student success. Our hope is that the early warning system, such as the one presented in this paper, will provide a means to maintain consistent throughput levels of undergraduate students while encouraging high enrollment numbers from learners with disadvantaged backgrounds. These vulnerable learners who may not be fully prepared or equipped to the culture of the university (i.e technical skills, adaptability to a new environment, and even to a different language for communication, English) can be supported by the higher-education institutions by responding to their needs in a fair manner [7].

This study uses a synthetic dataset by [5] to train several predictive models. These predictive models include support vector machines, random forests, coarse decision trees, extreme gradient boosted trees, linear logistic regressions, and C4.5 decision trees. We identify which classification model best deduces class label, which is one of the following four risk profiles: 'lowest risk,' 'medium risk', 'high risk', and 'highest risk'. Finally, we use the best model, with high predictive performance to deduce the most significant (important) factors in predicting the risk status. The information gain ranking algorithm will be used to rank the explanatory variables by their entropy to predict the class label.

The results show that student attrition or classifying the students into the right risk profiles is dominantly affected by biographical characteristics, followed by individual attributes. The pre-college characteristics show minimal or no effect on deducing student risk profiles. The overall performance of the models demonstrates that the fitted models perform well on an imbalanced class dataset; compared to models fitted with controlled balanced class dataset using the SMOTE algorithm. The random forest model achieved the best results with an accuracy of 85%, kappa statistic of 0.7, overall precision of 0.79, overall recall of 0.66, F-score of 0.69, and an AUC of 0.95 over the four risk profiles.

This research contributes to the field of educational data mining by providing a broader view of student attrition. The higher education institutions will have an indication of the factors that affect student attrition. The significance of this paper is to increase throughput rates at higher-education institutions by promoting student success initiatives.

II. Related Work

This section expands on the introductory background section that addresses the current literature of the stated problem. The study of student attrition dates back to the early 1900s by the researchers like [6] [8], till recent studies by [5], [9]; where the authors explored the factors affecting student academic performance. However, there is a growing demand for more advanced ways of analyzing educational data and incorporate more information.

A. Characteristics of the Learner

Student performance is an important metric used to track student and institutional goals, both long term, and short term educational goals. The progress of a student at a tertiary institution is determined by their course final mark or grade, which indicates progress to higher courses [10]. In higher education institutions, there are countless factors within and outside of school that affect the performance of students. The factors that came forward are socio-economic and psychological factors [11].

In recent years, many studies have focused on distinguishing the critical factors in the cluster of student factors that ensure success academically. Characteristics, for example, psychological wellness and social abilities, in particular, self-viability, inspiration, frames of mind and conduct, scholarly competence, communication abilities, team effort, participation, and group capacities, are among the significant highlights for the students to strive or cope at university [11]. Students with these aptitudes can work viably with others and deal with their studies productively [11] [12].

In this research, we adopt the conceptual framework model by [6], where he relates the background or family, individual attributes, and pre-schooling attributes, to the drop out decisions. These features are then used as input to predict student attrition. The combination and relation of these features influence the student's commitment to their goals and school.The impact of these attributes has been explored in previous studies, and provide a good prediction for student performance at higher education institutions.

Family or Background attributes explored by previous studies include: age, gender, race description, language, family background, living location, parent's occupation and qualifications to predict the student performance ([13] [5] [9] [14] [10] [8] [15] [16] [17] [18]). The findings are quite consistent- gender was found to be a high influencing factor in school drop out, with 68% probability [14]. Similarly, findings showed statistically significant gender differences and anxiety around academic achievement among South African university graduate students [8]. [19] [8] [10] the English language was found as one of the contributing factors to poor performance, among other factors.

In terms of pre-schooling attributes, quite an extensive research has been done on this section by assessing the student's summative assessments ([13] [14] [18] [10] [19] [15] [20] [21] [5] [9]). These factors include entry qualifications and the subjects taken by the student before college. Entry qualifications, and the pre-taken subjects before university show variability in the performance of the students [15]. A large portion of these research articles have concentrated on student performance in the U.S. and Europe regions. However, since social contrasts may play a role in forming the essential elements that influence student's performance, it is necessary to investigate features according to the region or country [11].

B. Predictive Modeling

Predictive modeling is the use of historical data to train the model, to discover patterns and behavior, then use that information to infer the likelihood of the class variable given the data. Many types of research have adopted the use of machine learning or data mining predictive models than traditional statistical models. This is due to the flexibility of machine learning models and their ability to incorporate vast and complex datasets.

In the field of education data mining, many authors have applied the use of predictive modeling like K-Nearest Neighbourhood (KNN) ([16] [22]); Decision Trees (DT) ([16] [23] [16] [17]); prediction of student grade ([13] [9]); and predicting performer or under performer [18].

Researchers like [23] extensively researched applying the Bayesian network and decision tree in the forecast of learner's academic behavior. The research reveals that the decision tree performed better than the Bayesian network. However, [21] discovers that when considering the rates of prediction, the applied data mining algorithm's performances are quite similar. The outcome of the research by [16] reveals that k-NN performed more than the other algorithms having a sensitivity of about 87%, decision tree also performed excellently with 79.7%, followed by Neural Network (NN) with 76.8%, while Naive Bayes got 73.9%; on the investigation of the dropout scenario of an online study platform. The Naive Bayes applied by [17] achieved excellent outcomes with a dropout precision of 0.917 and a recall of 0.924.

III. METHODOLOGY

In this paper we use the framework by [6] to predict student attrition by classifying the students into the four risk profiles. A student who enrolls in a higher-education program can fall into any of the four risk profiles that are associated with the probability of completing their program. The risk profile are defined as follows: ‘Lowest risk’ - where the student is expected to complete their degree in the minimum time (3 years); ‘medium risk’ - where the student is expected to complete in more than the minimum time; ‘high risk’ - where the student fails or drops-out before the minimum time; and ‘highest risk’ - where the student fails in more than the minimum time.

A. Data Collection and Pre-Processing

The data that was used in this study was synthetically generated by a Bayesian network which models the conceptual variables in [6] using cause-and-effect declared relationships. The details of the data construction can be found in [5]. The forward sampling algorithm was applied when generating the data. The values of the parent nodes are sampled from a conditional distribution, and the children-node values are sampled from their respective parent-sets. The sampling process follows a topological ordering and is iterative until all the node values are generated. Since the synthetic data was used, there was no ethical considerations to be made.

The synthetic dataset used in the study had 41 variables and 50 000 sampled observations. After data pruning and feature selection, variables were reduced down to 24. The features were selected by their relevance in terms of our aims and objectives. The main focus of this study is on biographical characteristics, pre-college observations, and enrollment observations in science degrees.

B. Features

In this study we adopt the conceptual framework by [6] to explain the assumption of causal relationships of the variables and the target variable, which is the risk status. The framework hypothetically assumes that the contributing factors to student performance are (i) Biographical characteristics, (ii) Pre-College Observations, and (iii) University Enrolment Observations. The Table I summarises the features used under each category.

TABLE I: The list of features generated by [5] used in the study categorised by the framework [6].

Biographical characteristics	**Pre-college**	**Enrollment observations**
Rural/Urban school Home Country Age at first year Home Province	School Quintile Mathematics Major English FAL Computers Additional Maths NBTAL, NBTMA, NBTQC	Year Started Plan Description Prob of (Math, Physics, Earth, Biological) Aggregate

For (i) Biographical characteristics, we used the following features: the age at first year, home province, home country - where the person originates from, and whether the person is from rural or urban school. For (ii) Pre-college observations we used the school quintile - which indicates the schooling poverty with quintile one being the poorest and quintile five as the least poor school, core mathematics, English first-additional language, computer studies, technical mathematics, and national benchmark tests (NBTAL, NBTMA, NBTQC) - which measure the student’s academic readiness for university. Finally, for (iii) Enrolment observations, we used the year the degree was started, plan description - the professional career, the Science streamline of the learner (mathematics, physical science, earth science, and biological science), aggregate for course marks.

In selecting the most important features, we used the Information Gain Ranking Algorithm (IG) to deduce each features entropy to predict the class variable. A higher IG, when compared to other features, indicates a higher importance of the feature for this prediction task. The IG scale ranges from zero to one, with zero being the least contributing and one being the most contributing.

C. Classification and Evaluation of Models

In this section, we discuss the predictive models that were trained by the data:

Random Forests (RF), also referred to as random decision forests, are an ensemble learning method for classification and regression. They fit several decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The random forest implemented by this paper followed that of [24]

Coarse Decision Trees are the simplest of decision trees. They provide very low model flexibility and only have a few leaves to make a coarse distinction between classes. The course decision tree implemented in this paper follows that of [25] and had 4 splits.

Linear Logistic Regression is of the most popular and simplest models used to model the linear relationship between the a dependent variable (Y) and independent variables (X_i), where i represents the feature. The ‘logistic’ refers to a categorical response variable; for two categories, it is a binary or dichotomous (binomial/binary logistic regression). However, It could have more than two classes (multinomial logistic regression).

The C4.5 Decision Trees is a supervised classification algorithm that is an extension of Quinlan’s earlier Iterative Dichotomiser 3 (ID3) algorithm used to generate a decision tree developed by [26].

Extreme Gradient Boosting, well known as (XGBoost) is an algorithm that makes use of gradient boosting decision tree algorithms. It computes residuals of prior fitted models, then uses these to create new models that will correct these errors, and thereby improving each new model, until they can no longer be improved. XGBoost applies the gradient descent algorithm to reduce the training error on new models [27].

Support Vector Machines (SVM), are a type of supervised learning algorithm that is applied to both regression and classification problems. They are usually applied to classification problems. SVM's create a linear line (hyperplane) that separates different (distinct) classes [28].

Model Evaluation: The following tools and metrics are used to evaluate the predictive accuracy of the trained models: confusion matrix, Kappa statistic, and area under the curve (AUC). A 10-fold cross validation will be used.

IV. Results and Discussion

In this section we present and discuss the results obtained from our experiment (study). The results represent the performance of the six fitted machine learning models in deducing the four risk profiles: 'lowest risk', 'medium risk', 'high risk', and 'highest risk'. The first section looks at the feature importance results and the second section is the classification models results.

A. Feature Importance

This section explores the contribution of each feature in classifying risk profile (status) using Information Gain (IG) or entropy. Table II shows the ranking of the features according to their contribution to classifying the risk profiles of the student. The first column (Rank), is the ranking of features from 1 to 24, most significantly contributing (high IG), to the least contributing (lowest IG). The second column is the feature name associated with the ranking. The last column represents the Information Gain (entropy), which is the value $0 \leq e \leq 1$, with 0 as no information gain, and 1 highest IG.

In Table II, the features are color-coded differently; biographical characteristics are light blue; pre-college observations are coded light purple; and individual characteristics are blank, is not shaded, as per [6] framework. The top 3 contributing features are (i) plan description, the student's career choice, (ii) the year started the program, which falls under the individual's characteristics, and (iii) the home province. The features ranked from 4 to 5 are (iv) home country, the students country of origin; and (v) the school description being either rural or urban. Feature rank 3 and 4 suggest that biographical characteristics are dominant in deducing the student risk status, followed by some few individual attributes. The pre-college attributes show no or minimal effect on student risk profiles. These results have tremendous implications for understanding the characteristics of learner attrition and more research is necessary to explore this.

B. Classification

In this section, we look at the results from the six fitted machine learning classification algorithms: random forests, linear logistic regression, coarse decision trees, extreme gradient boosted trees, support vector machines, and C4.5 decision trees. The different metrics that were used to evaluate the predictive performance of our models are confusion matrices, classification accuracy, Kappa statistic, sensitivity, Precision, F-1 score, and Area Under the Curve (AUC).

TABLE II: The Information Gain (entropy) ranking of features.

Rank	Feature	Information Gain
1	Plan Description	0.25
2	Year Started	0.24
3	Home Province	0.08
4	Home Country	0.04
5	Rural or Urban	0.02
6	Prob Of Math Streamline	0.00
7	Prob Of Physics Streamline	0.00
8	Prob of Earth Streamline	0.00
9	Prob of Biology Streamline	0.00
10	Aggregate	0.00
11	Number Of Years for Degree	0.00
12	Age at First Year	0.00
13	Quintile	0.00
14	Mathematics Matric Major	0.00
15	English HL	0.00
16	English FAL	0.00
17	Computers	0.00
18	Additional Mathematics	0.00
19	NBTAL	0.00
20	NBTMA	0.00
21	NBTQL	0.00

a) The accuracy: Accuracy was evaluated using 10-fold cross validation method. The classification accuracy is the metric used to measure how many predictions were correctly classified from all the predictions made. Figure 1 describes the results of the classification accuracy for the fitted models.

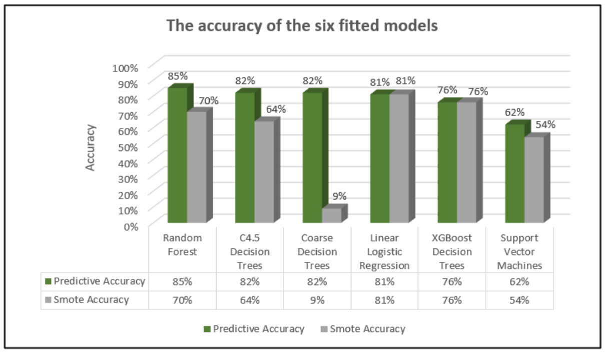


Fig. 1: The bar graph of the accuracy for the six fitted models.

The bars color coded green represent the predictive accuracy of the models trained with class imbalanced dataset, and the the bars color coded grey represent the smote accuracy, which is the accuracy obtained from the models trained with a corrected class imbalance dataset using SMOTE algorithm.

Comparing the green bars (predictive accuracy) and grey bars (smote accuracy), by model, we can see that the predictive accuracy for most models is higher than the smote accuracy except for the linear logistic regression and the XGBoost model. This can imply that the actual dataset with imbalanced classes, performs better or rather achieves greater predictive accuracy than the smote'd dataset with balanced classes. This is such contrary to improving class imbalance as it is a well known remedy for improving model performance. This phenomena can be explained by the process applied by the the smote algorithm when generating new points or synthetic examples, it does not take into consideration the neighbouring examples from other classes, which this then results in overlapping of classes and introduces noise in the dataset. This then results in poor performance of models in distinguishing the different classes.

b) Kappa statistic: The Kappa statistic is one of the most important metrics for measuring classifier performance, more particularly imbalanced class dataset. It is a good indicator of how the classifier performed across all classes because relying on the accuracy of the imbalanced skewed class dataset can give biased results. A kappa value of less than 0 means poor or no agreement, 0 means random agreement, and close to 1 means perfect agreement between predicted classifications and actual class labels [29].

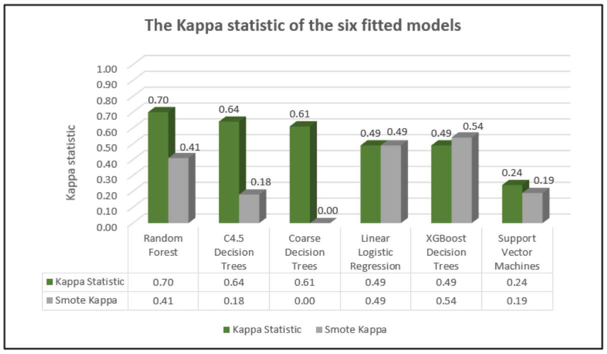


Fig. 2: The bar graph of the Kappa statistic for the six fitted models.

Figure 2 describes the kappa statistics from both the imbalanced dataset and the smote balanced dataset. The green bars represents the kappa statistic from models trained with imbalanced datasets, and the grey coded bars represent the kappa statistic from models trained with the class balanced dataset.

The Figure 2 is ordered by the kappa statistics obtained from the imbalanced dataset models in descending order. The random forest model is ranked number 1, which means it has the highest kappa value of 0.70, which is interpreted as substantial agreement according to [29] interpretation. When comparing the kappa statistic and smote kappa for the same model, random forest, the kappa value for the imbalanced dataset is higher than the one for the balanced class dataset. This also gives us an indication that the models trained with class imbalanced dataset performs better than models trained with class balanced dataset using smote algorithm.

c) Confusion Matrix: A confusion matrix is also known as the error matrix, which is a table used to summarize pre-known predicted labels. It is used to evaluate a classification model using the testing dataset. This metric is used to describe the confusion between the classes. The $n*n$ matrix; rows represent actual class labels, and the columns represent predicted class labels. The confusion matrices are computed from the models trained with the imbalanced class datasets because they achieved higher accuracy and higher kappa statistic, which makes them best compared to models trained with balanced class data using smote algorithm.

Figure 3 shows the confusion matrices for the six fitted models and their respective predictive performances.

d) The Sensitivity / Recall: The sensitivity or recall is the measure of the proportion of actual positives that are correctly classified. Table III illustrates the recall metric for the six trained models and each risk profile class. This will help us describe which models correctly classify risk profiles and which risk profile classes have a higher proportion of correctly classified labels.

The model that has the best (highest) overall recall is the random forest (0.66), followed by the C4.5 decision trees (0.61), then the coarse decision trees (0.53), then the linear logistic regression (0.51), then the xgboost decision tree (0.46), and the SVM has the least recall (0.35). This implies that the random forest has the highest proportion of correctly classified risk profile labels. The svm the least recall, meaning that they have a lower proportion of classes correctly classified as their actual label.

Looking at the recall rate by classes, shows that the high risk profile class has the highest recall, which means that most of the observations labeled at high risk class are actually high risk profiles; followed by medium risk profile; then highest risk profile; and lastly the lowest risk profile, meaning that the observations classified as lowest risk profiles, are actually not lowest risk label. The lowest risk profile class has the highest miss classification rate across all the models, and the high risk class has the most significant classification rate.

TABLE III: The Sensitivity (Recall) of the six trained models.

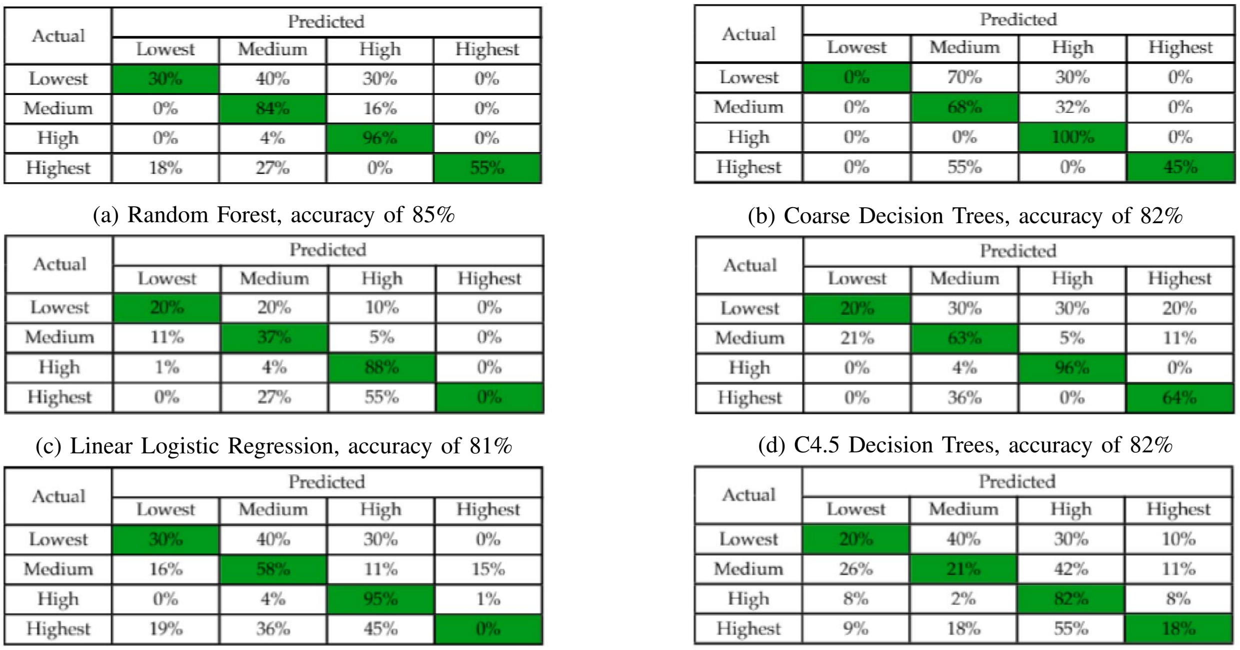
Model	Lowest	Medium	High	Highest
Random Forest	0.30	0.84	0.96	0.55
Linear Logistic Regression	0.40	0.70	0.95	0.00
Coarse Decision Trees	0.00	0.68	1.00	0.45
XGBoost Trees	0.30	0.58	0.95	0.00
Support Vector Machines	0.20	0.20	0.82	0.18
C4.5 Decision Trees	0.20	0.63	0.96	0.64

e) Precision: The precision refers to the percentage of the results which are relevant, meaning that if the model predicts a true class, how often is it correct? Table IV illustrates the results of the precision metric for the six fitted models, and the different class levels (risk profiles). The higher the precision value, the better the model is at predicting relevant risk profiles quite often.

TABLE IV: The Precision of the six trained models.

Model	Lowest	Medium	High	Highest
Random Forest	0.60	0.62	0.93	1.00
Linear Logistic Regression	0.40	0.47	0.90	NA
Coarse Decision Tree	NA	0.50	0.90	1.00
XGBoost Trees	0.38	0.50	0.89	0.00
Support Vector Machines	0.13	0.33	0.80	0.18
C4.5 Decision Trees	0.33	0.55	0.95	0.64

The random forest model has the highest overall precision (0.79), and the highest across all risk profile classes; followed



Actual	Predicted			
	Lowest	Medium	High	Highest
Lowest	30%	40%	30%	0%
Medium	0%	84%	16%	0%
High	0%	4%	96%	0%
Highest	18%	27%	0%	55%

(a) Random Forest, accuracy of 85%

Actual	Predicted			
	Lowest	Medium	High	Highest
Lowest	0%	70%	30%	0%
Medium	0%	68%	32%	0%
High	0%	0%	100%	0%
Highest	0%	55%	0%	45%

(b) Coarse Decision Trees, accuracy of 82%

Actual	Predicted			
	Lowest	Medium	High	Highest
Lowest	20%	20%	10%	0%
Medium	11%	37%	5%	0%
High	1%	4%	88%	0%
Highest	0%	27%	55%	0%

(c) Linear Logistic Regression, accuracy of 81%

Actual	Predicted			
	Lowest	Medium	High	Highest
Lowest	20%	30%	30%	20%
Medium	21%	63%	5%	11%
High	0%	4%	96%	0%
Highest	0%	36%	0%	64%

(d) C4.5 Decision Trees, accuracy of 82%

Actual	Predicted			
	Lowest	Medium	High	Highest
Lowest	30%	40%	30%	0%
Medium	16%	58%	11%	15%
High	0%	4%	95%	1%
Highest	19%	36%	45%	0%

(e) Xgboost Decision Trees, accuracy of 76%

Actual	Predicted			
	Lowest	Medium	High	Highest
Lowest	20%	40%	30%	10%
Medium	26%	21%	42%	11%
High	8%	2%	82%	8%
Highest	9%	18%	55%	18%

(f) Support Vector Machines, accuracy of 62%

Fig. 3: The confusion matrices showing the model performances of the six predictive classifier models on the test dataset.

by the C4.5 decision trees (0.62), then the coarse decision trees (0.60), then the linear logistic regression (0.44), then the xgboost (0.44), and the least is the svm (0.36). The svm has the least precision meaning that this models prediction's are often not correct, high miss classification rate.

The class with the highest precision is the high risk profile class, followed by the medium risk profile class, then the highest risk profile, and lastly, the lowest risk profile with the least precision, which means that most of the models are failing to predict the lowest risk profile class.

f) F-Score: The precision and recall are significant metric scores, but it is difficult to maximize both of them, so one has to trade off one metric. Taking the mean of these two metrics is misleading, because take for instance a classifier with a recall of 70%, and the precision of 10%; taking the average gives us an F score of 40%, whereas taking the weights into considerations will give us an F score of 18%, which is able to detect that our classifier is not doing well, which is where the concept of F score comes from or plays a role at. The F - score is a harmonic mean of precision and recall, that one can be able to find the right model that maximizes F1 score, and thereby maximizing both precision and recall.

Table V is the F score of the six fitted classification models.

TABLE V: The F-score of the six trained models.

Model	Lowest	Medium	High	Highest
Random Forest	0.40	0.71	0.95	0.71
Linear Logistic Regression	0.40	0.56	0.92	NA
Coarse Decision Tree	NA	0.58	0.95	0.63
XGBoost Trees	0.33	0.54	0.92	NA
Support Vector Machines	0.16	0.26	0.81	0.18
C4.5 Decision Trees	0.25	0.59	0.96	0.64

In the Table V, random forest has the highest F score when compared to the other five fitted models; and the svm has the lowest F score, which means on overall, the model is not performing well in classifying the risk profile classes.

g) Area Under the Curve: The area under the curve (AUC) represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at separating different risk profile classes. An AUC is a value between 0 and 1, where 1 means a good measure of separability; 0 means the separability is bad, and 0.5 means the model is not able to distinguish the different classes; it assigns labels at random.

Table VI describes the results of the AUC for each of the six trained models.

TABLE VI: The Area Under the Curve (AUC) of the six fitted models.

Model	AUC
Random Forest	0.95
XGBoost Trees	0.92
Coarse Decision Tree	0.91
C4.5 Decision Trees	0.91
Linear Logistic Regression	0.83
Support Vector Machines	0.77

The first column in Table VI is the model name, and the second column is the AUC value in descending order. The top-ranked model has the highest AUC when compared to the other models, which is the random forest with AUC of 0.95; followed by the xgboost decision tree with AUC of 0.92; then the coarse decision tree with AUC of 0.91; then the C4.5 decision trees with AUC of 0.91; then the linear logistic classifier with AUC of 0.83; and the model with the least AUC is the SVM with AUC of 0.77.

This means that the random forest has the highest measure of separability, and the svm has the least capability of distinguishing risk profile classes.

V. Discussion and Conclusion

Our study of deducing the students into the correct risk profiles using biographical (background), individual, and schooling characteristics showed that student attrition or classifying the students into the right risk profiles is dominantly affected by biographical characteristics, followed by individual attributes. The pre-college characteristics show minimal or no effect on deducing student risk profiles. Similar results were achieved by [5], in that the eight most significant (contributing) attributes are biographical and individual characteristics. They play a major (important) role in deducing the student into the correct risk profiles, as per the [6] conceptual model.

The observations from the results in general, demonstrate that the fitted models performed well on an imbalanced class dataset; compared to models fitted with controlled balanced class dataset using SMOTE algorithm. This can be due to the process applied by the smote algorithm while generating synthetic samples, it does not take into consideration the neighbouring samples from other classes. It can then result in overlapping of classes and can introduce additional noise.

The fitted machine learning algorithms were successfully able to detect (deduce) the different risk profiles. However, the positive class detection rate differs for different class proportions, i.e., majority class, high risk profile has higher rates of positive (correct) detection of this class; compared to the minority classes, lowest and highest risk profiles, have a high negative rate (miss classification); across all the models. The impact of skewed class sizes of the training and testing set would need to be investigated further on the classification accuracy.

The random forest model achieved the best results with an accuracy of 85%, kappa statistic of 0.7, and an AUC of 0.95 over the four risk profiles. The accuracy and kappa statistic were used to select the best model parameters. Even though there were misclassifications that were noted in the confusion matrices, Figure 3, they were not severe. For example the random forest confused 40% of lowest risk students as medium risk, which is not severe compared to 27% of highest risk students classified as medium risk. Misclassifying lowest risk as medium risk is less of an error than classifying medium risk as highest risk, or vice versa.

This research contributes to an argument for a more complex view of predicting undergraduate student attrition by including the student's biographical, individual, and schooling characteristics.

The study of student attrition is one of the most important studies in the educational sciences. Future work could include the use of real data which would provide us with an opportunity to verify if our theoretical model is applicable in the real-world scenario, or can distinguish between the different risk profiles. It will also shed light on the most important features or characteristics that affect student attrition.

Currently, our method used the duration it took to complete the degree (qualification), to derive the target variable- risk profile; for example, the longer it takes to finish the degree, the higher the risk profile. The future study will use the target variable attrition (dropout) or not, and apply binary classification modeling, then use the model to get predicted probabilities of attrition; then calibrate (bin) these probabilities of attrition; for example: $0\% - 25\%$ - bin 1, and is the lowest risk profile; $26\% - 50\%$ - bin 2, medium risk profile; $51\% - 75\%$ - bin 3, high risk profile; and $76\% - 100\%$ - bin 4, highest risk profile. Then evaluate how well the model is able to deduce students into the correct risk profiles. The study concludes that student attrition is affected by biographical and individual attributes, and deploying tailor made interventions in a timely manner will help mitigate some of the challenges that contribute to student attrition.

Acknowledgment

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

References

[1] J. Kuzilek, M. Hlosta, D. Herrmannova, Z. Zdrahal, and A. Wolff, "Ou analyse: analysing at-risk students at the open university," *Learning Analytics Review*, pp. 1–16, 2015.

[2] R. A. Johnson, R. Gong, S. Greatorex-Voith, A. Anand, and A. Fritzler, "A data-driven framework for identifying high school students at risk of not graduating on time," in *Bloomberg Data for Good Exchange Conf*, vol. 5, 2015.

[3] M. Dell and S. Dell, "A brief guide to recreational pyromania," Available at http:// impact.msdf.org/university-completion/ (2019/08/05).

[4] P. Anand, A. Herrington, and S. Agostinho, "Constructivist-based learning using location-aware mobile technology: an exploratory study," in *EdMedia+ Innovate Learning*. Association for the Advancement of Computing in Education (AACE), 2008, pp. 2312–2316.

[5] T. Abed, R. Ajoodha, and A. Jadhav, "A prediction model to improve student placement at a south african higher education institution," in *2020 International SAUPEC/RobMech/PRASA Conference*. IEEE, 2020, pp. 1–6.

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

[7] M. Cross and C. Carpentier, "'new students' in south african higher education: institutional culture, student performance and the challenge of democratisation," *Perspectives in Education*, vol. 27, no. 1, pp. 6–18, 2009.

[8] T. S. Mwamwenda, "Gender differences in scores on test anxiety and academic achievement among south african university graduate students," *South African Journal of Psychology*, vol. 24, no. 4, pp. 228–230, 1994.

[9] R. Ajoodha, A. Jadhav, and S. Dukhan, "Forecasting learner attrition for student success at a south african university," in *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.* ACM, 2020.

[10] C. T. Downs *et al.*, "Investigation into the academic performance of students in bioscience at the university of natal, pietermaritzburg, with a particular reference to the science foundation programme students," 2002.

[11] S. T. Hijazi and S. Naqvi, "Factors affecting students' performance." *Bangladesh e-journal of sociology*, vol. 3, no. 1, 2006.

[12] E. Lust and F. C. Moore, "Emotional intelligence instruction in a pharmacy communications course," *American journal of pharmaceutical education*, vol. 70, no. 1, p. 06, 2006.

[13] M. M. Abu Tair and A. M. El-Halees, “Mining educational data to improve students’ performance: a case study,” *Mining educational data to improve students’ performance: a case study*, vol. 2, no. 2, 2012.

[14] S. Pal, “Mining educational data using classification to decrease dropout rate of students,” *arXiv preprint arXiv:1206.3078*, 2012.

[15] E. Alfan and N. Othman, “Undergraduate students’ performance: the case of university of malaya,” *Quality assurance in education*, vol. 13, no. 4, pp. 329–343, 2005.

[16] E. Yukselturk, S. Ozekes, and Y. K. Türel, “Predicting dropout student: an application of data mining methods in an online education program,” *European Journal of Open, Distance and e-learning*, vol. 17, no. 1, pp. 118–133, 2014.

[17] B. K. Bhardwaj and S. Pal, “Data mining: A prediction for performance improvement using classification,” *arXiv preprint arXiv:1201.3418*, 2012.

[18] U. K. Pandey and S. Pal, “Data mining: A prediction of performer or underperformer using classification,” *arXiv preprint arXiv:1104.4163*, 2011.

[19] L. Steenkamp, R. Baard, and B. Frick, “Factors influencing success in first-year accounting at a south african university: A comparison between lecturers’ assumptions and students’ perceptions,” *South African Journal of Accounting Research*, vol. 23, no. 1, pp. 113–140, 2009.

[20] E. Alfan and N. Othman, “Undergraduate students’ performance: the case of university of malaya,” *Quality Assurance in Education*, vol. 13, no. 4, pp. 329–343, 2005. [Online]. Available: https://doi.org/10.1108/09684880510626593

[21] D. Kabakchieva, “Predicting student performance by using data mining methods for classification,” *Cybernetics and information technologies*, vol. 13, no. 1, pp. 61–72, 2013.

[22] M. Mayilvaganan and D. Kalpanadevi, “Comparison of classification techniques for predicting the performance of students academic environment,” in *2014 International Conference on Communication and Network Technologies*. IEEE, 2014, pp. 113–118.

[23] N. T. Nghe, P. Janecek, and P. Haddawy, “A comparative analysis of techniques for predicting academic performance,” in *2007 37th annual frontiers in education conference-global engineering: knowledge without borders, opportunities without passports*. IEEE, 2007, pp. T2G–7.

[24] M. Pal, “Random forest classifier for remote sensing classification,” *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[25] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[26] J. R. Quinlan, “Combining instance-based and model-based learning,” in *Proceedings of the tenth international conference on machine learning*, 1993, pp. 236–243.

[27] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, “Xgboost: extreme gradient boosting,” *R package version 0.4-2*, pp. 1–4, 2015.

[28] R. G. Brereton and G. R. Lloyd, “Support vector machines for classification and regression,” *Analyst*, vol. 135, no. 2, pp. 230–267, 2010.

[29] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *biometrics*, pp. 159–174, 1977.