Using Ensemble Decision Tree Model to Predict Student Dropout in Computing Science

Mohammed Naseem
School of Computing, Information and Mathematical Sciences
The University of the South Pacific
Suva, Fiji
mohammed.naseem@usp.ac.fj

Kaylash Chaudhary
School of Computing, Information and Mathematical Sciences
The University of the South Pacific
Suva, Fiji
kaylash.chaudhary@usp.ac.fj

Bibhya Sharma
School of Computing, Information and Mathematical Sciences
The University of the South Pacific
Suva, Fiji
bibhya.sharma@usp.ac.fj

Aman Goel Lal
School of Computing, Information and Mathematical Sciences
The University of the South Pacific
Suva, Fiji
goel.lal@usp.ac.fj

***Abstract*— Science, Technology, Engineering and Mathematics (STEM) professionals play a key role in the development of an economy. STEM workers are critical thinkers as they contribute immensely by driving innovations. There is a high demand for professionals in the STEM fields but there is also a shortage of human resource in these areas. One way to reduce this problem is by identifying students who are at-risk of dropping out and then intervening with focused strategies that will ensure that these students remain in same the programme till graduation. Therefore, this research aims to use a data mining classification technique to identify students who are at-risk of dropping out from their Computing Science (CS) degree programmes. The Random Forest (RF) decision tree algorithm is used to learn patterns from historical data about first-year undergraduate CS students who are enrolled in a tertiary institute in the South Pacific. A number of factors are used which comprise of students demographic information, previous education background, financial information as well as data about students' academic interaction. Feature selection is performed to determine which factors have greater influence in students' decision in dropping out. Cross-validation techniques are used to ensure that the models are not over-fitted. Two models were built using a 5-fold and 10-fold cross-validation and the results were compared using several measures of model performance. The results show that the factors corresponding to students' academic performance in a first-year programming course had the greatest impact student attrition in CS.**

Keywords*— *STEM, Computing Science, drop out, data mining, feature selection, cross-validation

I. Introduction

Developing countries in the Pacific region are undergoing an information and communication technology (ICT) revolution. This digital transformation has empowered Pacific Island Countries (PICs) to overcome the issue of geographical remoteness allowing them to be connected to the global economy [1]. The recent increase in mobile phone usage and improved Internet penetration has enabled better connectivity thus providing new employment and business opportunities [2]. Governments of the PICs are focusing on leveraging ICT to generate new jobs for their citizens through e-governance and e-services. As a result, the governments are moving towards building a knowledge-based society that would become vital for diminishing poverty.

This has led to an increase in the demand for higher education as more and more students are now trying to attain a degree in their respective fields of study. Although the labor market in the informal employment sector is more prevalent than the formal sector, there is still a need for professionals in the latter. In the 21st century, with the rise in demand for technological innovation, CS is a powerful educational tool for fostering critical thinking, problem solving, and creativity. CS provides students with computational literacy and problem-solving skills that are desperately needed by the workforce [3]. More importantly, CS is a field that ensures that students are competitive and adaptable in the labor market, not just for jobs in CS, but for many occupations that increasingly require computer skills.

The CS field is one of the fastest growing and highest paying career paths in the world. The information age has steered a lot of demand for CS jobs. Despite these compelling facts, the number of students graduating from CS is relatively lower compared to the number of students admitted in the programme. Student attrition, a critical determinant of the success of higher education (HE), still remains a pressing issue for the institutes. High attrition rates for CS students is a serious problem that must be addressed by HE if the need for technologically competent professionals is to be met. There is a need to identify factors that contribute to the poor retention of students in CS, in other words, coerce student to dropout.

This study is inspired by the need to identify the reasons of students drop out from CS programmes in the developing PICs. In past, many studies have been conducted that focus on student attrition but little work has been carried out in identifying reasons that lead to student dropout in CS programme. Nevertheless, while these studies focused on student attrition in developed countries, little have been done on CS programme student dropout in the developing countries. The findings of this study are expected to shed light on factors that contribute to dropout rates in the developing countries.

II. Literature Review

A. Student Attrition Models

Student attrition is a key issue for many HEIs and has been a challenge for many researchers over the past decades. Early researches in determining university student dropout were based on surveys and questionnaires. Statistical methods were predominantly used as the tool for analyzing the data collected from the surveys and questionnaires which were used in constructing theoretical models to address the student attrition problem. The theoretical models aimed to understand the reasons behind student dropout.

The dropout model by Tinto in 1975 [4] is considered as the groundwork for the continuous research in the field of student attrition and is considered as the most widely accepted theoretical model. It is an interactive model that explains the reasons behind student withdrawal process. The two key variables of the interactive model are goal commitment and institutional commitment. Tinto states that as academic and social integration increases, the students' commitment towards their educational goals also increases, and ultimately their decision to continue their study. The Undergraduate Dropout Process model by Spady in 1971 [5] is another popular theoretical model which was based on two main institutional systems: the academic system and the social system. The assumption is that the student attrition is determined by the interaction between the student and the academic environment in which the student's characteristics are influenced by a variety of factors.

A major drawback of these models is that they are not very effective at predicting students who are at risk of dropping out and even if identification is possible it is often too late to avoid student dropout [6]. Recently, there is an increase in the use of data mining techniques to predict student dropout from HE. Classification is a data mining technique used by many researchers studying student attrition phenomena. Classification algorithms come under supervised learning technique which uses the training data sample to learn patterns and construct models that can predict the class label of the testing data samples.

Delen (2012) [7] used several classifiers to build analytical models to predict freshmen student attrition in a public university in the mid-west region of the United States. Decision Trees (DT), Artificial Neural Networks (ANN) and Logistic Regression (LR) were used to train and test the predictive models. In a similar study by Pal (2012) [8], a predictive model was built using the DT algorithm to predict student dropout in an Engineering program to identify students who need support from the student dropout program. The classification models built included the ID3, C4.5, CART and ADT decision tree algorithms that used five years of data collected from the student's management system of an engineering institute in India. Using the DT algorithm to identify students that have high risk of attrition is the main contribution in [9]. The authors use the ID3 algorithm to develop a predictive model using student admission data of students doing Bachelor of Technology from an Engineering college. In another study by Ghadeer and Alaa (2015) [10], several data mining classifiers were used to examine and predict dropouts for CS students at Al-Aqsa University. The DT and NB classification algorithms were built to predict student dropout which used data consisting 1290 records of CS students. Using one year of freshmen undergraduate student data, Oztekin (2016) [11] applied three data mining models to predict degree completion of students from a public university in USA. The classifiers used as prediction algorithms were DT, ANN and SVM. Orozco and Niguidula (2017) [12] applied DT, NB and Rule Induction classification models to predict student dropout after first semester in their freshmen year. In a more recent study, Bayesian networks were used in [6] to predict the likelihood of students abandoning their CS degree in a university in Spain. The Bayesian networks used by the authors as classifiers were Naïve Bayes, TAN, K2 and PC.

Table I describes some classification algorithms used in the literature to predict student attrition, the performance measures used the respective accuracies.

B. Student Attrition Factors

In the past, researchers have studied a variety of factors that contribute to student attrition. These factors are categorized into demographic variables, prior educational background, academic performance, and financial background. Demographic factors refer to human characteristics such as age, gender, parental education, ethnicity, finances, and health. These demographic characteristics have been found to be the major contributing variables to student attrition in HE by various researchers

TABLE I. COMPARISION OF CLASSIFICATION MODEL ACCURACIES USED IN STUDENT DROPOUT PREDICTION IN CS

Author	DM Classification Models	Performance Metrics
Lacave et al. [6]	Bayesian Networks: • NB • TAN • **K2 (best log-likelihood score)** • PC	Log likelihood score for 5-fold and 10-fold cross-validation
Orozco & Niguidula [12]	• **NB (83.50%)** • DT (82.96%) • Rule Induction (82.65%)	Models were trained and tested using 10-fold cross-validation. Accuracy was the main metric for evaluation of model performance.
Delen [7]	• **ANN (81%)** • DT (78%) • LR (74%)	Models were trained and tested using 10-fold cross-validation. Accuracy was the main metric for evaluation of model performance. False Positive and False negative rates were also considered.
Pal [8]	Decision Trees: • **ID3 (85.7%)** • C4.5 (80.8%) • CART (67.7%) • ADT (72.4%)	Precision scores using 10-fold cross-validation technique
Arora & Badal [9]	• DT	None (The DT algorithm was the only technique used to build a predictive model in this study)
Ghadeer & Alaa [10]	• **DT (98.14%)** • NB (96.86%)	Models were trained and tested using 10-fold cross-validation. Accuracy was the main metric for evaluation of model performance.
Oztekin [11]	• DT (73.75%) • ANN (71.59%) • **SVM (77.61%)**	10-fold cross-validation was used. Measures of model performance were accuracy, sensitivity and specificity

[12, 7, 8, 10, 11]. On the contrary some researchers argue that demographic attributes such as gender [6] and ethnicity [13] do not correlate with student persistence.

A growing body of literature has studied academic factors as key predictive variables of student dropout in HE. The determinants of dropout rate include; first-year academic performance, class attendance, academic satisfaction, subjects passed and student engagement [6, 7, 8, 11]. The strongest consensus on the key determinant of student retention/attrition is student performance. According to Tinto [4], student retention or continuance in HE is largely dependent on student interactions with the institution's academic and social systems, rather than a function of student or institutional variables. The greater the degree of student integration into the institution, the greater will be the institutional commitment thus student's commitment to completing HE.

Prior educational performance is also a much-debated determinant of student dropout in HE. Several scholars are affirmative of high school performance as indicators of persistence [6, 14, 7, 8] while other researchers [15] argue otherwise; that exam grades do not have a positive correlation to student's decision to attrite.

Numerous studies have been conducted over the years to debate on financial factors as a predictive feature of student attrition. There are varied discussions on parental income or financial aid influences on student persistence. Even though some researchers [7, 8] have seen a correlation between student persistence and the ability to pay for study expenses, most other researchers [11, 4] are assertive that financial aid has little influence on student persistence. These authors argue that financial factors are secondary determinants of student dropout.

C. Contributions

Although there has been a number of studies focusing on predicting student attrition, the authors do not unanimously agree on the factors that are influential in students' decision to drop out [16]. Additionally, very few studies have considered identifying factors that contribute to student drop out from CS programmes. This paper makes the following contribution:

- A wide spectrum of features are utilised in the prediction of student attrition in CS programmes including demographics, academic, financial and prior education history. A unique subset of features comprising of data about students online presence in a course is explored in this study,
- An ensemble classification technique is used to a build predictive model to identify students at the risk of dropping out from CS degree programme.
- Several measures of model performance are used to determine the best model that correctly identifies students at the risk of dropping out from CS programmes.

III. Research Methodology

The CRISP-DM (Cross Industry Standard Process for Data Mining) framework for data mining applications was used in this research. Fig. 1 illustrates the six steps of CRISP-DM, which were used to explore the patterns in the institutional data used in this work [7].

A. Business Understading

Issues in higher education were identified while conducting a literature survey. It was observed that high rates of student attrition in CS degree programmes is a major problem faced by HEIs in developing countries including tertiary institutes in the small South Pacific Island nations. While getting higher admission numbers is often preferred, retaining existing students is considered a cheaper option. Early identification of students at-risk of dropping out from CS programmes can assist in improving university retention rates. Thus, the use of predictive modelling can enable faculties to determine students who are more likely to withdraw and then intervening with focused assistive strategies which will ensure that these students remain in the programme until graduation. Three research questions were developed considering the findings of the literature survey:

i. What factors are likely to predict freshmen student dropout in a CS degree programme at a university?

ii. Which students are most likely to drop out from their first-year CS degree programmes?

iii. Can ensemble techniques such as the RF decision tree models predict student attrition in CS with high accuracies?

B. Data Understading

This step requires data collection and familiarisation. The features that are used by data scientists and researchers studying freshmen student attrition were identified.

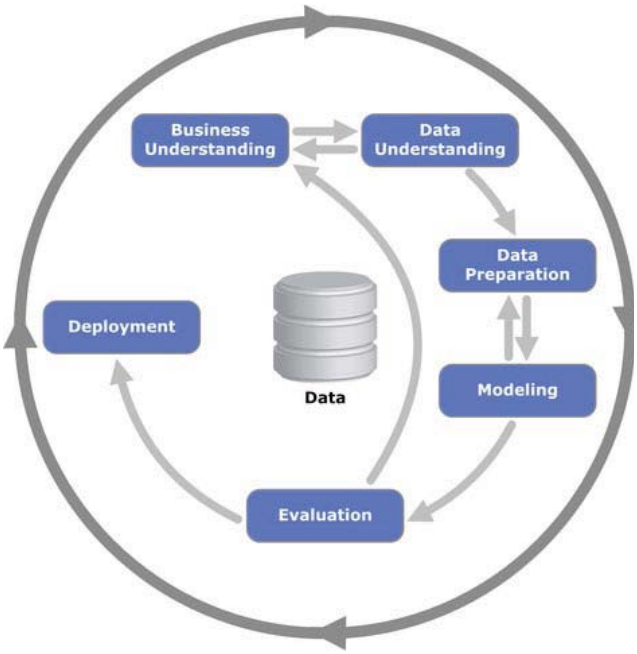

Fig. 1. CRISP-DM Model
Source: Oztekin [11]

The features can be categorised as demographic factors, prior education background, financial factors and academic performance. The data used in this study was collected from a single institution in the South Pacific region. Past five years of historic data from 2013 to 2017 about students doing CS degree at The University of the South Pacific (USP) was used.

Data was obtained from various data sources and integrated into a single flat file. Student's enrollment data was collected from the University Student Academic Services (SAS) database that contained demographic information of the students as well as the prior education background. Additionally, the course interaction and assessment information was gathered from Moodle which is the Learning Management System used by the University.

C. Data Preparation

In the data preparation step, data cleaning and feature selection takes place. Real world data is usually noisy as it contains missing values and outliers, which must be treated. Data cleaning is an important step in data mining and most data scientists consider this step as most time-consuming as it takes about 80% of the time [17]. If noise is not removed from the final data used in prediction then it will certainly affect the performance accuracies.

Intially, data was collected from various data sources so it was obtained in different comma-seperated (.csv) format. All data was then integrated into a single Microsoft Excel file. Missing values for continous variables were replaced with mean values while the label, "unknown" was used for missing categorical variables. Feature scaling is a data pre-processing method where the range of the independent variables are normalised. In this step, the min-max technique was used to normalise the range of the variables to [0,1]. The formula for min-max normalization is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

where x is the original value and x' is the normalized value.

New features were derived by combining attributes and performing calculations. For example, students age at enrollment was calculated using the date of birth and enrolled year attributes. The final dataset consisted of 963 observations and 23 attributes. The features are described in Table II.

An analysis of various demographic characteristics of the selected sample is presented in Fig. 2. The distribution of students amongst various campuses of the university shows that majority of the CS students were from the main campus (Laucala) constituting of 67.5% (650/ 963) compared to 32.5% (313/963) who studied in other regional campuses.

Majority of the students majoring in CS were from the Bachelor of Science (BSC) programme contributing a total of 77.7% (748/963) followed by 6.33% (61/963) from Bachelor of Arts (BA) and 4% (39/963) from Bachelor of Commerce (BCOM). The remaining 12% of students were from other programmes. It is also interesting to note that 63.9% (615/963) of the first-year CS students were part-time and the remaining 36.1% (348/963) studied full-time. Science, Technology, Engineering and Mathematics (STEM) courses have an underrepresentation of women worldwide thus the gender distribution in the sample was expected. Male students made up 74.7% (719/963) while the females constituted a small proportion of 25.3% (244/963). There was a huge disparity in the marital status of the CS students. A dominant 89.9% were single students while a very small representation of students were married.

D. Modelling

The tool used for the data mining model creation is the open-source statistical software, R that is widely used among statisticians and data miners.

TABLE II. DESCRIPTION OF FEATURES USED

Attribute Name	Description
CAMP	Campus where the student had studied. Possible values: ***{Laucala, Regional}***
PROG	Student's programme of study. Possible values: ***{BA, BSC, BCOM, Other}***
AGE	Student's age at enrollment. Continuous variable ranging from 16-47.
STYPE	Possible values: ***{Part-time (PT), full-time (FT)}***
GENDER	Student's Gender ***{F, M}***
MSTATUS	Whether the student was single or married during enrollment ***{S, M}***
DISAB	Whether the student had any disability. This is a binary variable ***{0, 1}***
NATION	Student's nationality
HSL	Highest Secondary School level ***{YR12, YR13, Foundation, Other}***
SPON	Whether the student studied privately or under sponsorship ***{0, 1}***
HMG	Student's high school math grade ***{A+, A, B+, B, C+, C,D,E,EX,N}***
A1SCORE	First assignment marks in CS111
A2SCORE	Second assignment marks in CS111
ASUB	Number of assignments submitted in CS111
QATTEMPT	Number of quizzes attempted in CS111
FPOSTS	Number of Moodle forum posts in CS111
CPACCESS	Frequency of Moodle Course page access in CS111
AQM	Average quiz marks in CS111
GRADE	Highest Grade in CS111
S1UNITS	Number of courses taken in first semester of study
INCOUNTRY	Whether the student studied in the country of residence ***{0, 1}***
INTC	Whether the student studied in the town/city of residence ***{0, 1}***
DROPOUT	This is a dichotomous variable where the positive class indicated by binary variable "1" corresponding to first-year CS students who have dropped out from their programme and the negative class "1" corresponding to those who returned in the following year and continuing in the same programme as the previous year.

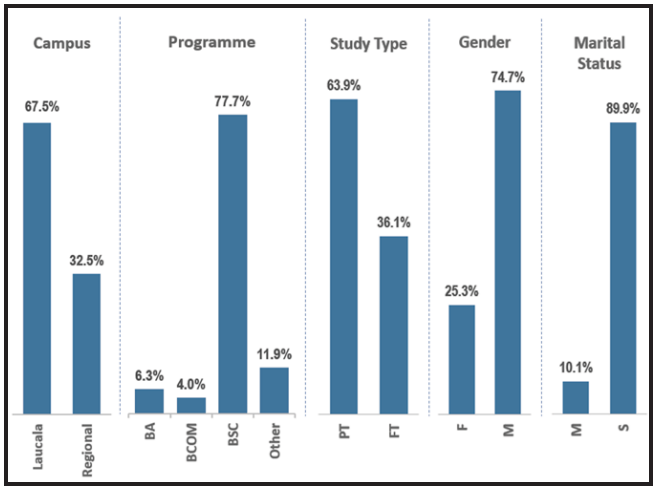

Fig. 2. Statisitcs for demographic variables

R is considered a powerful machine learning tool due to the breadth of techniques it offers for data analysis, visualisation, sampling, supervised learning and model evaluation [18]. It is free software that is increasingly becoming popular in the area of data science [19].

Since this is a classification problem, the data mining classifier algorithm selected was RF. RF is based on decision tree (DT) algorithm which uses Information Gain to decide which attribute must be placed in the root node. The root node is further divided into leaf nodes using a threshold and then the steps are repeated to identify a variable-threshold pair that maximises the homogeneity of the resulting two or more subgroups of samples. The result is a single decision tree. However, unlike the DT algorithm that only constructs a single tree, the RF model is an ensemble technique which uses multiple decision trees in order to predict.

E. Evaluation

The dataset was partitioned into train and test sets with an 80-20% split. K-fold cross-validation technique was used in the experiment to test the ability of the predictive model on unseen data [7]. The test set was used to validate the results of the model. A number of performance measures were used to determine how accurately the model was able to identify students who are at the risk of dropping out from their CS degree programmes.

Accuracy is the measure of overall predictive precision, which is the ability of the model to differentiate the dropout and non-dropout cases correctly. The proportion of true positive and true negative in all evaluated cases gives the accuracy of the model. The formula to calculate the accuracy is represented as

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

where:

- True Positive (**TP**) = the number of cases correctly identified as non-dropout
- False Positive (**FP**) = the number of cases incorrectly identified as non-dropout
- True Negative (**TN**) = the number of cases correctly identified as dropout
- False Negative (**FN**) = the number of cases incorrectly identified as dropout

The ***sensitivity*** of a model is its ability to determine the non-dropout cases correctly. It is the proportion of true positive in non-dropout cases. The formula to calculate the sensitivity is represented as:

$$\textbf{Sensitivity} = \frac{TP}{TP+FN} \quad (3)$$

The ***specificity*** of a model is its ability to determine the dropout cases correctly. It is the proportion of true negative in dropout cases. The formula to calculate the specificity is represented as:

$$\textbf{Specificity} = \frac{TN}{TN+FP} \quad (4)$$

The receiver operating cost (ROC) curve is a method of comparing the different models graphically where the sensitivity is plotted against 1-specificity. The area under the curve (AUC) is another measure of model performance, where the greater the AUC, the better the model performance.

F. Deployment

This step involves the implementation of the data mining predictive models in the business to solve real-world problems. The implementation of the models is beyond the scope of this research but it would be considered for future researches.

IV. Results

Statistical analysis was performed on the final dataset containing all features. A correlation analysis using the Spearman rho correlation test was performed. Fig. 3 shows the correlation plot of the attributes. The size and color of the circles represent the strength of correlation the attributes have between each other. The blue circles represent a positive correlation whereas the orange ones indicate a negative correlation. The attributes that have a whitespace indicates that no significant correlation existed between the variables. It is important to identify the correlation of each attribute with the class variable, which in this case was DROPOUT. The attributes that had no significant correlation with the class attribute has a greater chance of being excluded from the subset of important features. These are attributes corresponding to DISAB, PROG, INTC, SPON and GENDER.

Table III shows the results of the feature selection algorithm. The Boruta algorithm in R was used to identify attributes which have significant contribution in students' decision to drop out from their CS programmes. The Boruta algorithm rejected five out of 23 attributes and these are the same variables that were found to have no significant correlation with the class variable.

After the important features were identified, then two predictive models were trained and tested using the RF algorithm. Different cross-validation techniques were used to build the models where the first model (Model 1)

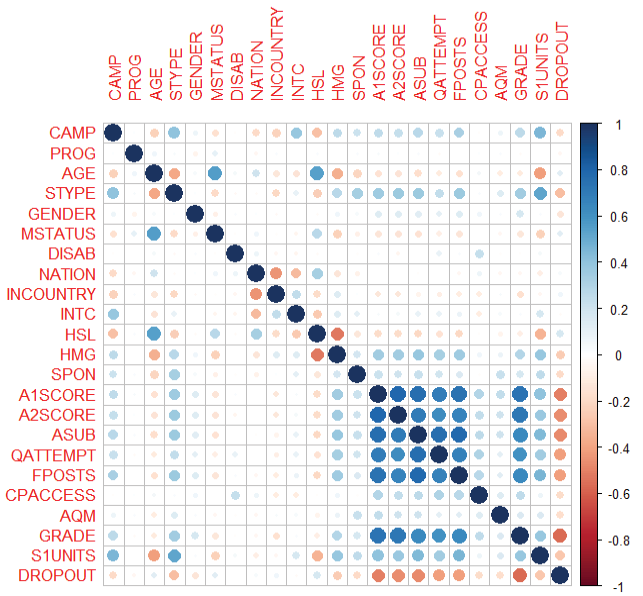

Fig. 3. Correlation plot for the variables

used a 5-fold cross-validation and the second model (Model 2) used a 10-fold cross-validation. The confusion matrix was used to evaluate the performance of these models.

Table IV provides the data about the various performance metrics used to evaluate the performance of the two RF predictive models. The accuracy of both the models were greater than 80%, which is comparable with the results from the literature [7, 8, 11, 12]. Model 1 had a sensitivity score of 73.68% which was higher than the sensitivity score of Model 2 suggesting that Model 1 was able to identify the dropout cases more accurately than the Model 2. Both models predicted the non-dropout labels with the same accuracy. The Kappa score of Model 1 was significantly higher than that of Model 2. This means that instances in Model 1 were more accurately predicted when compared to the expected values.

The ROC curve illustrated in Fig. 4 also shows that the RF model based on 5-fold cross-validation performed more accurately when compared to the other model. Another measure to confirm that the performance of the first model was significantly better than the performance of the second model are the respective AUC scores provided in Table V. The AUC for Model 1 was 0.8629, which was higher than the AUC of Model 2, 0.8550. All the performance measures are in favor of the model which was trained using the 5-fold cross-validation.

V. Discussion

In this paper, a predictive model was built using an ensemble decision tree model known as the RF algorithm to determine which first-year undergraduate CS students would drop out from their programmes. A number of factors that lead to student dropout in CS were identified as important that are consistent with the existing literature. These factors include students' academic performance such as assessment marks and grades in the first-year programming course, the workload in the first semester of study, demographic factors including students' age, marital status, nation, campus, study type and whether the student studied in his/her own country and students previous education background.

Factors corresponding to students' performance in first-year programming course have been found to be the strongest predictors of drop out in this study. This suggests that there is a need to identify the reasons for the poor performance in first-year programming course in the PICs. One way to mitigate the risk of poor performance is the use of early warning systems that would alert students about the adverse outcome early enough for them to improve [15]. This will also allow instructors to plan early intervention strategies to assist students who are at the risk of failing the course [20]. Mobile devices can be actively integrated to student learning activities to garner better engagement and interaction which can contribute to better performances. However, student readiness, perceptions and attitude towards these new devices have to be considered before their implementation [21, 22, 23].

Programming is perceived as a difficult field of study by many students. Past studies have shown a significant correlation between mathematics ability and students' performance in programming [24, 25]. Therefore, there should be an early detection of mathematics ability of the students prior to entering CS degree programmes. Through the use of tools such as the Online Mathematics Diagnostic Test (OMDT) [26], the numeracy gaps can be identified and students who do not meet the required level for CS can be recommended to additional support mechanisms to bridge those gaps at an early stage. Moreover, students who have work and family commitments are often unable to attend face-to-face sessions which also affects their performance.

TABLE III. Attribute Importance Table

Attribute	MeanImp	Decision
DISAB	-2.0002314	Rejected
PROG	0.1389159	Rejected
INTC	0.3633271	Rejected
SPON	1.9392175	Rejected
GENDER	2.3603811	Rejected
MSTATUS	**4.0124384**	**Confirmed**
AGE	**4.7474221**	**Confirmed**
NATION	**5.2622254**	**Confirmed**
CAMP	**5.9364784**	**Confirmed**
INCOUNTRY	**7.0055878**	**Confirmed**
HMG	**7.0376998**	**Confirmed**
HSL	**7.2692499**	**Confirmed**
CPACCESS	**7.3076194**	**Confirmed**
S1UNITS	**7.94948**	**Confirmed**
FPOSTS	**8.1452247**	**Confirmed**
AQM	**10.6750424**	**Confirmed**
ASUB	**11.6967507**	**Confirmed**
A2SCORE	**13.585312**	**Confirmed**
STYPE	**15.4730476**	**Confirmed**
QATTEMPT	**17.4215541**	**Confirmed**
A1SCORE	**21.2154627**	**Confirmed**
GRADE	**40.5403609**	**Confirmed**

TABLE IV. RANDOM FOREST MODEL PERFORMANCE

Performance measures	*Model 1 5-fold*	*Model 2 10-fold*
Accuracy	0.8177	0.8011
Senstivity	0.7368	0.6974
Specificity	0.8762	0.8762
Kappa	0.6209	0.5842

Students who miss out on the concepts introduce in lectures are often found to disengage from the courses. Therefore, it is important that universities invest in lecture capture systems which will allow students to access lecture recordings

VI. CONCLUSIONS

A predictive model is presented in this paper that was built using the RF algorithm to identify which CS students in their first year of studies have a likelihood of dropping out. The model built using the 5-fold cross-validation was found to be more accurate at predicting drop out cases. Early identification of students' intention to drop out would enable universities to intervene with focused strategies at individual level to ensure that these students remain in the programme. Academic performance in the first-year programming course was found to be the strongest predictor of drop out. Therefore, this study suggests that it is imperative for universities to identify the reasons students in the PICs find difficulty in programming courses. Once the reasons for poor performance are known then the instructors of these courses can utilise mitigating techniques to assist the students to attain a successful grade.

Future recommendations are to use data at granular level to gain better insights about student dropout in CS. Data quality is essential for obtaining accurate results in learning analytics.

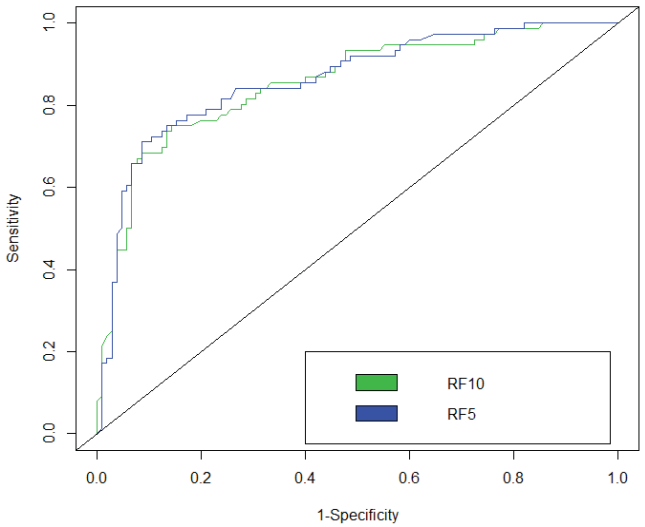

Fig. 4. ROC curve for the predictive models

TABLE V. AREA UNDER THE CURVE

	AUC Scores
Model 1 (RF5)	0.8629
Model 2 (RF10)	0.8550

For example, it was worth noting that students' financial information was not considered to be an important predictor of dropout. This could be due to the fact that very less information was known about the type of sponsorship. Therefore, it would be interesting to find whether the type of sponsorship has any correlation with student dropout. Furthermore, only a single data mining classification algorithm was used in this study so future work will focus on use of more classifiers such as Support Vector Machines and Naïve Bayes.

REFERENCES

[1] E. Reddy and B. Sharma, "Mobile Learning Perception and Attitude of Secondary School Students in the Pacific Islands.," in *Proceedings of the 22nd Pacific Asia Conference on Information Systems (PACIS 2018)*, Yokohama, Japan., 2018.

[2] M. Minges and C. Stork, "Economic and social impact of ICT in the Pacific.," Sydney, Australia, 2015.

[3] M. M. Mhashi. and A. L. I. A. Alakeel, "Difficulties facing students in learning computer programming skills at Tabuk University.," in *Proceedings of the 12th International Conference on Education and Educational Technology (EDU'13)*, Iwate, Japan, 2013.

[4] V. Tinto, "Dropout from Higher Education: A theatrical synthesis of recent research.," *Review of Education Research,* vol. 45, pp. 89-125, 1975.

[5] W. G. Spady, "Dropouts from higher education: An interdisciplinary review and synthesis," *Interchange,* vol. 1, no. 1, pp. 64-85, 1970.

[6] C. Lacave, A. I. Molina and J. A. Cruz-Lemus, "Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks," in *Behaviour & Information Technology*, 2018.

[7] D. Delen, "Predicting Student Attrition with Data Mining Methods," *Journal of College Student Retention: Research, Theory & Practice,* vol. 13, no. 1, pp. 17-35, 2012.

[8] S. Pal, "Mining Educational Data to Reduce Dropout Rates of Engineering Students," *International Journal of Information Engineering and Electronic Business,* vol. 2, pp. 1-7, 2012.

[9] R. K. Arora and D. Badal, "Predicting Students Attrition using Data Mining," *International Journal of Computer Science & Engineering Technology (IJCSET),* vol. 4, pp. 1338-1341, October 2013.

[10] A.-O. S. Ghadeer and E.-H. M. Alaa, "Data Mining In Higher Education: University Student Dropout Case Study," *International Journal of Data Mining & Knowledge Management Process (IJDKP),* vol. 5, no. 1, pp. 15-27, 2015.

[11] A. Oztekin, "A hybrid data analytic approach to predict college graduation status and its determinative factors," *Industrial Management & Data Systems,* vol. 116, no. 8, pp. 1678-1699, 2016.

[12] M. E. Orozco and J. C. Niguidula, "Predicting Student Attrition Using Data Mining Predictive Models," in *Proceedings of 143rd The IIER International*

Conference, Jeju Island, South Korea, 2017.

[13] A. W. Astin, Predicting academic performance in college: Selectivity data for 2300 American colleges., New York, NY: US: Free Press., 1971.

[14] 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.

[15] A. Harvey and M. Luckman, "Beyond demographics: Predicting student attrition within the Bachelor of Arts degree.," *International Journal of the First Year in Higher Education,* vol. 5, no. 1, pp. 19-29, 2014.

[16] S. d. O. Durso and J. V. A. d. Cunha, "Determinant factors for undergraduate student's dropout in an accounting studies department of a Brazilian public university.," *Educação em Revista,* vol. 34, 2018.

[17] R. M. Isiaka, R. D. Babatunde, F. J. Ajao and S. O. Abdulsalam, "A Machine Learning Approach to Dropout Early Warning System Modeling," *International Journal of Advanced Studies in Computers,* vol. 8, no. 2, pp. 1-12, 2019.

[18] J. Brownlee, "Machine Learning Mastery," Machine Learning Mastery Pty. Ltd., 15 January 2016. [Online]. Available: https://machinelearningmastery.com/use-r-for-machine-learning/. [Accessed 1 October 2019].

[19] R. Gour, "Towards Data Science: Sharing concepts, ideas, and codes.," 26 April 2019. [Online]. Available: https://towardsdatascience.com/a-complete-guide-to-learn-r-29e691c61d1. [Accessed 1 October 2019].

[20] A. Jokhan, . B. Sharma and S. Singh, "Early warning system as a predictor for student performance in higher education blended courses," *Studies in Higher Education,* 2018.

[21] E. Reddy, P. Reddy, B. Sharma, K. Reddy and M. G. M. Khan, "Student Readiness and Perception to the Use of Smart Phones for Higher Education in the Pacific," in *Proceedings of Asia-Pacific World Congress on Computer Science and Engineering*, Nadi, Fiji, 2016.

[22] P. Reddy and B. Sharma, "Effectiveness of Tablet Learning in Online Courses at University of the South Pacific," in *Proceedings of Asia-Pacific World Congress on Computer Science and Engineering*, Fiji, 2015.

[23] B. Sharma, A. Jokhan, R. Kumar, R. Finiasi, S. Chand and V. Rao, "Use of Short Message Service for Learning and Student Support in the Pacific Region," in *Y. Zhang, Handbook of Mobile Teaching and Learning*, Springer, 2015.

[24] J. Owolabi, P. Olanipekun and J. Iwerima, "Mathematics Ability and Anxiety, Computer and Programming Anxieties, Age and Gender as Determinants of Achievement in Basic Programming," *GSTF International Journal on Computing (JoC),* vol. 3, no. 4, pp. 109-114, 2014.

[25] I. L. Balmes, "Correlation of mathematical ability and programming ability of the computer science students.," *Asia Pacific Journal of Education, Arts and Sciences,* vol. 4, no. 3, pp. 85-88, 2017.

[26] B. N. Sharma, R. Nand, M. Naseem, E. Reddy, S. Narayan and K. Reddy, "Smart Learning in the Pacific: Design of New Pedagogical Tools," in *Proceedings of 2018 IEEE International Conference of Teaching, Assessment, and Learning for Engineering (TALE)*, NSW, Australia, 2019.