Education and Information Technologies
https://doi.org/10.1007/s10639-022-11018-3

Predicting Freshmen Attrition in Computing Science using Data Mining

Mohammed Naseem[1] · **Kaylash Chaudhary**[1] · **Bibhya Sharma**[1]

Received: 25 August 2021 / Accepted: 22 March 2022

Abstract
The need for a knowledge-based society has perpetuated an increasing demand for higher education around the globe. Recently, there has been an increase in the demand for Computer Science professionals due to the rise in the use of ICT in the business, health and education sector. The enrollment numbers in Computer Science undergraduate programmes are usually high, but unfortunately, many of these students drop out from or abscond these programmes, leading to a shortage of Computer Science professionals in the job market. One way to diminish if not completely eradicate this problem is to identify students who are at risk of dropping out and provide them with special intervention programmes that will help them to remain in their programmes till graduation. In this paper, data mining techniques were used to build predictive models that can identify student dropout in Computer Science programmes, more specifically focusing on freshmen attrition since a significant number of dropout occurs in the first year of university studies. The predictive models were built for three stages of the first academic year using five classification algorithms which were Random Forest, Decision Tree, Naïve Bayes, Logistic Regression, and K-Nearest Neighbour. The models used past five years of institutional data stored in university's repositories. Results show that the Naïve Bayes model performed better in stage 1 with an AUC of 0.6132 but in stages 2 and 3, the overall performance of the Logsitic Regression models were better with an AUC of 0.7523 and 0.8902, respectively.

✉ Mohammed Naseem
mohammed.naseem@usp.ac.fj

Extended author information available on the last page of the article

Published online: 04 April 2022

Keywords Attrition · Dropout · Retention · At-risk students · Freshmen · Data mining

1 Introduction

Student attrition in higher education (HE) is a social problem that affects every tertiary institute (Lacave et al., 2018). It is an alarming concern adversely impacting higher education institutes (HEI) as high levels of attrition continues to hinder the economic status, reputation and long term planning of colleges and universities (Delen, 2012). Students' success is a key indicator of an institution's overall performance (Murtaugh et al., 1999) and an important determinant of government grants. According to Evans (2000), high levels of attrition decreases graduation rates, consequently damaging the reputation of the university, thus increasing attrition further. High dropout rates are not only costly for HEIs, but these also create a feeling of inadequacy in students and may lead to one being socially stigmatised (Lacave et al., 2018; Schneider et al., 2018). On the other hand, student attrition leads to subsequent loss of potential skills and knowledge that would have translated into future workforce contributing to a nation's economic growth.

The Computer and Information Technology (IT) industries are booming worldwide. Developing countries are experiencing a digital transformation which is driven by the massive increase in mobile phone usage and Internet penetration (Reddy & Sharma, 2018). In the Pacific, which is sparsely populated amongst many small islands, the use of mobile phones and affordable Internet prices has enabled people to get connected easily thereby addressing many economic challenges and problems associated with geographic remoteness (Sharma, et al., 2015, 2017). It has also led to an ease of access to a number of services such as health, finance and education (Reddy & Sharma, 2015, 2018), which also contributes to the economic growth in the region. This Information and Communication Technology (ICT) revolution in the Pacific has compelled many organisations to leverage technology in their business processes to enhance productivity and efficiency. Consequently, creating a greater demand for more qualified personnel in the IT sector. The impact of broadband penetration on employment growth range from 0.2% to 5.3% for every one percent increase in penetration. An increase of 20% by 2025 is approximated in ICT related jobs while new opportunities for programmers with better salaries are to be expected (Richards & Terkanian, 2013).

Notwithstanding the increasing demand for Computer Science (CS) professionals, there still exists a great shortage of employees with a Bachelors or higher degree in most of the IT occupations (Kori, et al., 2015). The ICT skills shortage is a real issue that is directly affecting the economy of small island nations in the Pacific. A potential solution to this problem is to find ways to decrease the rate of dropout of students from CS degree programmes. Retaining existing students is also less expensive than enrolling new ones (Shilbayeh & Abonamah, 2021). Hence, there is a genuine need for early identification of students who are at the risk of dropping out from the programme.

The use of modern technology has enabled storage of large volumes of data in different data sources. Discovering underlying patterns in these datasets is essential for the successful existence of any organization. The rapid expansion of data storage has led to new opportunities in terms of using analytical methods to identify knowledge and address challenges (Lacave et al., 2018; Badr et al., 2016; Yukselturk et al., 2014). Data mining is an analytical tool that has been used for various tasks including, market-basket analysis (Kaur & Kang, 2016), reducing customer churn (Dolatabadi & Keynia, 2017), customer relation management (Kazemi et al., 2015) and customer segmentation (Kansal et al., 2018), to name a few.

Data mining in education, more commonly known as Educational Data Mining (EDM), is a process of detecting hidden patterns from educational data using several data mining techniques (Baker & Kalina, 2009; Zaffar et al., 2018; Orozco & Niguidula, 2017). The various applications of EDM include prediction of student performance and progress (Badr et al., 2016; Patil et al., 2018), course recommendation (Al-Badarenah & Alsakran, 2016), identification of learning patterns for implementation of personalized learning (Lin et al., 2013) and prediction of student attrition (Lacave et al., 2018; Schneider et al. 2018; Delen, 2012; Pal, 2012; Ghadeer & Alaa, 2015; Oztekin, 2016; Gairín, et al., 2014).

Although there is significant literature on addressing student attrition at university and course level, very limited work has been done in the identification of at-risk students in Computing Science discipline. Furthermore, the main contribution in this work is the use of students' online presence attributes as predictors of student attrition. Moreover, the researchers do not always agree with the important predictors of student attrition because it is considered to be greatly dependant on the contextual elements (Olaya et al., 2020). Thus, the main objective of this research is to use predictive modelling to determine which students will attrite from their undergraduate CS degrees, particularly in their first year of studies at a regional university. The findings of this research will help universities to provide students with proper intervention programs such as personalized counselling, mentoring programmes, financial assiatance and corrective remedies so that they complete their CS degrees.

The remainder of the paper is structured as follows: discussion of related work from the literature is provided in Sect. 2. Section 3 outlines the methodology used in detail followed by the presentation of results in Sect. 4. In Sect. 5, a discussion of the implications of the results is provided and finally Sect. 6 concludes this work.

2 Background and Related Work

2.1 Student attrition models

Student attrition happens when a student either voluntarily or non-voluntarily drops out or withdraws from a course or a programme before the expected time of completion (Delen, 2012). Student attrition is a major problem faced by many HEIs globally and researchers have been investigating this challenge over the years in the pursuit to improve student retention. Early studies in the area of student attrition involved

the use of surveys and questionnaires where statistical methods were utilized as the main tool for analysing the data. Results obtained from these experiments were used to develop theoretical models that gave insights to the student attrition problem. The newer approach to address the student attrition issue is the use of predictive models. The next section highlights some theoretical models and various predictive models build using machine learning and data mining algorithms.

Tinto's (1975) *'Theory of student departure'* is a ground breaking principle for the continuous research in this field and it is also the most widely accepted theoretical model. Students' personal goal commitment and institutional commitment are the two key variables in this interactive model which demonstrated that as students get more integrated into the academic and social aspect of an institute their commitment to continue in the programme increases. The *'Undergraduate Dropout Process Model'* developed by Spady (1970) which is based on two main institutional systems: the academic system and the social system. According to Spady, the interaction between the student and the academic system plays an important role in student attrition. Spady's model emphasizes that students characteristics are influenced by many factors such as academic potentials, family background, academic achievements, peer support, faculty interaction, and the degree to which the student is socially integrated with the academic environment. These models have greatly influenced the selection of array of factors for the prediction of student attrition in this study. Although these models are considered to provide good insights about student attrition, they are not very effective in determining which students have the greater likelihood of dropping out. Even if identification of potential students was possible with these models, it would often be very late to prevent students from dropping out (Lacave et al., 2018).

Recently, there is an increase in technical approaches to addressing this problem. Due to the massive amount of data available about students through their interaction and engagement with the technological tools, many researchers and data scientists have started to invest in utilizing data mining techniques to uncover hidden knowledge that can be useful in reducing student attrition. Researchers have also used data mining techniques to build student attrition predictive models.

Most of the studies in the field of student attrition prediction differ in terms of the use of machine learning algorithms used. Classification techniques have been most commonly used in performing data mining tasks to predict student attrition. Delen (2012) used three classification algorithms to build predictive models that could identify freshmen students who are likely to drop out from a public university in the mid-west region of the United States. These models were built using Artificial Neural Network (ANN), Decision Tree (DT), and Logistic Regression (LR) where the ANN model had the best accuracy of 81%. The author used tenfold cross validation to train and test the models. Similarly, Pal (2012) used several DT classifiers to develop a student attrition predictive model using five years of data of students enrolled in an Engineering programme. ID3, C4.5, CART and ADT decision tree algorithms were used with tenfold cross validation but the ID3 model generated the best precision score of 85.7%. In another study by Ghadeer and Alaa (2015), several data mining classifiers were used to examine and predict dropouts for CS students at Al.-Aqsa University. The DT and Naïve Bayes (NB) classification algorithms

were built to predict student dropout which used data consisting 1290 records of CS students. Likewise, Oztekin (2016) used three classification algorithms which were DT, ANN, and Support Vector Machine (SVM) to predict degree completion of students in a public university in USA. In a similar study, Orozco and Niguidula (2017) built predictive models to identify freshmen student attrition. The models were built using DT, Naïve Bayes (NB), and Rule Induction (RI) classifiers to determine which students are the risk of dropping out after their first semester. The NB model outperformed its counterparts with an accuracy of 83.5%.

More recently, Kemper et al. (2020) used Decision Trees and Logistic Regression to predict student attrition at the Karlsruhe Institute of Technology. Their models were mostly based on examination data but also comprised some personal data that are easily available such as age and gender. The data was divided into three sets, each set containing only data available after the respective semester. Due to the underrepresentation of the minority class, the authors applied SMOTE balancing to resample their data, consequently getting six datasets in total. Their results showed that classification accuracy of at least 83% was already achievable after the first semester. Although classification techniques are predominant in predicting student attrition, some authors such as Shilbayeh and Abonamah (2021) have employed association rule mining to detect students who are at the risk of attrition. The authors used the Apriori algorithm to generate several association rules to identify potential students who would dropout from their MBA program. More sophisticated machine learning techniques have also been applied in the field of student retention and attrition. For example, Uliyan et al. (2021) investigated the students retention risk using deep learning techniques. The authors used the bidirectional long short term model (BLSTM) and the condition random field (CRF) methods to predict each student label independently in which they achived high level of accuracy. Interestingly, Olaya et al. (2020) implemented uplift modelling to address the student attrition problem. Uplift modelling, in contrast to conventional modelling, aims to estimate the effect of a treatment on the behaviour of an individual rather than targeting on the basis of risk. Hence, the uplift models are able to improve retention by targeting retention efforts to students who have a better likelihood of being retained.

While most of these studies have predominantly focussed on freshmen student attrition, some researchers have also attempted to predict student attrition in later years of the study period. For example, Yu et al. (2010) used machine learning to build predictive models for the sophomore year using classification trees, multivariate adaptive regression splines (MARS), and neural networks. Similarly, Blekic et al. (2017) examined the effects of factors related to second-year study in the prediction of student drop out from an institute by their third year. Due to the context of this study, the freshmen cohort was selected because the authors believed that majority of the students drop out after their first year of studies due to the challenges they face during this phase of their learning journey. Meanwhile, Yaacob et al. (2020) used three years of undergraduate student data to predict student dropout from Computer Science.

Furthermore, these predictive models have been built on different levels of study. Some models were built to predict student attrition at programme level (Pal, 2012; Rovira et al., 2017; Pérez et al., 2018) while some focussed on student dropout at

course level (Kovacic, 2010; Costa et al., 2017; Aguiar et al., 2014). Programme level student attrition aim to identify students who are at risk of dropping out of a particular programme, either to join another programme of study or completely abandon their higher education. On the other hand, course level student dropout deals with students leaving a course and not returning to complete it. This study is similar to the work of Ghadeer and Alaa (2015), Lacave et al. (2018), and Yaacob et al. (2020) in terms of predicting student attrition at discipline level, in particular, the Computing Science discipline. These studies aimed to identify factors that contributed to students' decision to abscond the Computing Science discipline and enrol in another discipline. This study differs from the work of Ghadeer and Ala in the sense that these authors used first two years of transcript data in their prediction while this study utilizes data related to first-year of studies only. The data used does not only refer to the transcript data but other features such as prior education background, demographics, financial factors, and students online presence data. Additionally, in the work of Lacave et al. (2018), the authors are only using Bayesian networks to build the predictive models and they used a smaller dataset of size 383 records. However, this study uses several other classifiers and the dataset is comparatively larger with 963 records.

Table 1 describes some classification algorithms used in the literature to predict student attrition, the performance measures used and the respective accuracies.

2.2 Student attrition factors

In the past, researchers have studied a plethora of factors that contribute to student attrition. These factors are categorised into demographic variables, prior educational background, academic performance, and financial background. While factors related to students' academic performance have often been found to be the leading variables in the prediction of student attrition, there is a need to explore the impact of other student factors on their decision to drop out.

Demographic factors refer to human characteristics such as age, gender, parental education, ethnicity, finances, and health. These factors are often used in the research related to student attrition because they are readily available. Most academic institutes gather these information about their students during the time of admission. Students' gender has been studied rigorously by most researchers and many have found gender differences to play a major role in students' decision to drop out (Ghadeer & Alaa, 2015; Pal, 2012; Yasmin, 2013). Another demographic variable used frequently is ethnicity which may refer to students' race, colour or even the geographic setting to which the student belongs (such as Polynesian, Melanesian, or Micronesian). Both Kovacic (2010) and Thammasiri et al. (2014) showed that there is a significant association between the ethnic background and student attrition. Age is another demographic factor that greatly influences students' decision to withdraw (Lacave et al., 2018). Apart from these other demographic variables that have been studied in the past include marital status (Pérez et al., 2018; Thammasiri et al., 2014; Yasmin, 2013), resident status (Aulck, et al., 2017; Pal, 2012; Thammasiri et al., 2014), employment status (Lacave et al., 2018; Kovacic, 2010; Yukselturk et al.,

Table 1 Comparision of classification model accuracies used in student dropout prediction

Author	DM Classification Models	Performance Metrics	Best Model
Lacave et al. (2018)	Bayesian Networks: -NB -TAN -**K2** -PC	Log likelihood score for **fivefold** and **tenfold** cross-validation	The K2 model had the best performance with highest log-likelihood score for both fivefold (-12.93) and tenfold (-12.99)
Orozco and Niguidula (2017)	-NB (83.50%) -DT (82.96%) -Rule Induction (82.65%)	Models were trained and tested using **tenfold** cross-validation. Accuracy was the main metric for evaluation of model performance	NB
Delen (2012)	-ANN (81%) -DT (78%) -LR (74%)	Models were trained and tested using **tenfold** cross-validation. Accuracy was the main metric for evaluation of model performance. False Positive and False negative rates were also considered	ANN
Pal (2012)	Decision Trees: -ID3 (85.7%) -C4.5 (80.8%) -CART (67.7%) -ADT (72.4%)	Precision scores using **tenfold** cross-validation technique	ID3 decision tree model had the best precision score 0f 85.7%
Ghadeer and Alaa (2015)	-DT (98.14%) -NB (96.86%)	Models were trained and tested using **tenfold** cross-validation. Accuracy was the main metric for evaluation of model performance	DT
Oztekin (2016)	-DT (73.75%) -ANN (71.59%) -SVM (77.61%)	**tenfold** cross-validation was used. Measures of model performance were accuracy, sensitivity and specificity	SVM

2014), disability (Kovacic, 2010) and region/state (Ghadeer & Alaa, 2015; Lacave et al., 2018; Pal, 2012; Thammasiri et al., 2014).

Students' academic performance has been prominently used in studies related to student attrition. Aulck et al. (2017) determined that grades received in the first year courses, particularly in math courses, were amongst top ten predictors of students' attrition. Similarly, Rovira et al. (2017) used grades for each first year course in Computing Science, Mathematics and Law programmes to successfully forecast which students would return to enrol in their second and third year. Grade Point Average (GPA) has also been used frequently as an academic performance indicator in predicting student attrition. Feature selection performed by Orozco and Niguidula (2017) revealed students grade average as a strong predictor of student attrition. Thammasiri et al. (2014) used several features to predict freshmen attrition and found Fall GPA to be one of the top ten predictors.

Some authors have also explored the effect of other academic related factors on student dropout. Kovacic (2010) found significant correlation between student attrition and course programme, and student attrition and semester in which the students studied their programme. Lacave et al. (2018) discovered that the highest course a student is enrolled in was an important predictor of student attrition. Thammasiri et al. (2014) identified that freshmen students who declared their programme major had a better likelihood of persisting in their programme.

Prior education background has also been studied by many researchers in the field of student attrition. A significant number of scholars are affirmative of high school performance, such as high school math grade, as an indicator of student attrition (Delen, 2012; Pal, 2012; Thammasiri et al., 2014; Dekker et al., 2009).

Student financial background as a predictor of student attrition has been a much debated topic over the years. Delen (2012) and Thammasiri et al. (2014) had investigated several factors related to students' finances which included received Fall financial aid, received Spring financial aid, Fall student loan, Fall grant/tuition waiver scholarship, Fall federal work-study, Spring student loan, Spring grant/tuition waiver scholarship, and Spring federal work-study. An important variable analysis was performed by the authors which showed that parents' financial background and type of financial aid were strong determinants of freshmen student attrition. Another study by Oztekin (2016) revealed financial support as a strong predictor of attrition. Their results showed that students who studied under scholarships were less likely to drop out compared to those who were studying on loan.

Apart from these factors, this study has also investigated the impact of features related to students' online presence in a course. These features are derived from students' interactions and engagement with the Learning Management System (LMS), in this case, Moodle.

2.3 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. Additionally, very few studies have considered identifying factors that contribute to student disengagement from their CS programmes. This paper makes the following contribution:

- A subset of features comprising of data about students online presence in a course is explored in this study,
- Another important contibution of this study is the building of data mining models at three different stages of the first academic year which include:
  - Stage 1: Pre-enrollment (before the student starts a course)
  - Stage 2: After the first semester (after the results of the first semester are available)
  - Stage 3: After the second semester (after the results of the second semester are available)

The effect of various student attrition factors can differ at different stages of learning and this is the rationale for building predictive models at the three different stages. For example, before the student makes any engagement with the course and when only the demographic, financial, and prior education background information is available, the main predictors will be different compared to when features related to students academic performance are included. Having the knowledge of which factors are more influential will enable decision makers to plan proper interventions at early stages to avoid attrition.

3 Methodology

The Cross-Industry Process for Data Mining (CRISP-DM) which is known to be a robust and well-proven methodology providing a structured approach to planning and data mining was used in this project. Using a standard methodology such as the CRISP-DM ensures quality in the data science projects. The CRISP-DM methodology has six major steps which are outlined below. The various phases of CRISP-DM interact with each other iteratively as illustrated in Fig. 1.

1. **Business understanding**—The first step of CRISP-DM involves data mining problem definition.
2. **Data understanding**—This step requires data collection and familiarization.
3. **Data preparation**—In this step, data cleaning and feature selection take place.
4. **Modelling**—This step involves the implementation of data mining algorithms to perform predictive analytics.

Fig. 1 CRISP-DM Model

5. **Evaluation**—This step involves analyzing the performance of the model(s) and reviewing the steps applied in the construction of the model to ensure that the business requirements were met.
6. **Deployment**—This step involves the implementation of data mining predictive models in the business to solve real-world problems.

3.1 Data

The data used in this study was from a single HEI from the South Pacific region which is the only regional campus in the South Pacific. The data consisted of demographic, financial, academic performance and online presence information of students enrolled in first year degree programmes at the university between 2013 and 2017. Data was acquired from several university databases and consolidated.

3.2 Data preparation

Data preparation is the most time-consuming step of data mining as it usually takes about 80% of the time to get the final dataset ready for modelling. The attrition data provided by the university did not include specific information about student attrition from a discipline, so specific cases had to be articulated to figure out which students had dropped out from their CS programmes. The following two cases were considered:

Case 1: Students who either transferred to another university or abandoned HE.

- If a student has not returned in the following year and he/she did not graduate then the student has dropped out from CS.

Case 2: Students who changed their majors.

- For a returning student who has not graduated, if the student has not changed his/her majors (that is, if he/she is no longer enrolled in the CS discipline) then the student has dropped out from CS.

3.2.1 Data integration

Data integration involves merging data gathered from different sources. Initially, data was collected from various databases, so it was obtained in different formats and as separate files. Individual files were integrated into a single flat file using the VLOOKUP function in Microsoft Excel 2016. Students' identification numbers were used as the primary key to merge the data from different files.

3.2.2 Data transformation

There was a need to aggregate three attributes in the data. Data aggregation involves the computation of new attributes by summarizing information from multiple records in the table. Some attributes such as the number of assignments in a course were not consistent in different semesters, so there was a need for aggregation. For example, the average assignment and average quiz marks were calculated using the individual assignment and quiz marks. New features were derived by combining attributes and performing calculations. For example, students' age at enrolment was calculated using the date of birth and enrolled year attributes. Feature scaling or normalization of continuous variables was done as well. Feature scaling is a data pre-processing method where the range of the continuous independent variables are normalized. The min–max technique was used to normalize the range of the variables to [0, 1]. The formula for min–max normalization is given as:

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

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

3.2.3 Missing values

Some attributes contained missing values which needed to be treated. Although there are several tools and packages that allow automatic treatment of missing values, it is preferred to use manual procedure as it helps in gaining a better understanding of the

data and allows to use the domain knowledge to treat the missing values more efficiently. The following techniques were used to replace the missing values:

- The attributes containing missing numerical values were replaced with average/mean value. Continuous attributes such as assessment marks and course work contained some missing values. Careful replacement of these missing values was done using the mean scores identified for different clusters which were determined using the course grade information.
- The attributes containing missing categorical values were replaced with "unknown". These instances were not excluded from the final dataset as they significantly contribute to the model creation and there would be loss of important knowledge if this information is ignored.

Three datasets were created for the predictive modelling of student attrition at three different stages of the first academic year. A description of the three models and the respective datasets is provided in Table 2.

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 (Luacala) 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.

3.2.4 Dataset balancing

The class distribution of the target variable for the three datasets is presented in Figs. 3, 4 and 5. The percentage of dropout cases for each dataset are 42%, 32% and 16%, respectively. It can be noted that the three datasets consisted of imbalanced class where the percentage of the dropout cases are relatively lower compared to the non-dropout cases. Class imbalance is a common problem in the predictive modelling of student attrition (Márquez-Vera et al., 2013; Rovira et al., 2017) as the number of students dropping out is often fewer than the number of students retained.

The class imbalance problem occurs when there is a disparity in the distribution of the target class (Thammasiri et al., 2014). For a binary target variable in an unbalanced dataset, one class label will have a higher frequency (referred to as the majority class) while the other class label will have relatively fewer frequency (referred

Table 2 Description of the dataset for the three predictive models

Stage	Description	Dataset Size	Features	Type
Stage 1 Model	Before the students starts a course	948	**CAMP** {0 = Regional, 1 = Laucala}	Nominal
			PROG {1 = BSC, 2 = BA, 3 = BCOM, 4 = OTHER}	Nominal
			AGE {> = 16}	Numeric
			STYPE {1 = FT, 2 = PT}	Nominal
			GEN {1 = Male, 0 = Female}	Nominal
			MSTAT {0 = Unknown, 1 = Married, 2 = Single}	Nominal
			DIS {1 = Yes, 0 = No}	Nominal
			NATION {1 = Fiji, 2 = SI, 3 = Vanuatu, 4 = Samoa, 5 = Tonga, 6 = Kiribati, 7 = Tuvalu, 8 = FSM, 9 = Other}	Nominal
			INCOUNT {1 = Yes, 0 = No}	Nominal
			INTC {1 = Yes, 0 = No}	Nominal
			HSECLEV {1 = Foundation, 2 = Year13, 3 = Year12, 4 = Other}	Nominal
			MGRADE {0 = N, 1 = E, 2 = D, 3 = C, 4 = C+, 5 = B, 6 = B+, 7 = A, 8 = A+}	Nominal
			SPON {1 = Sponsored, 0 = Private}	Nominal
			DROPOUT {1 = Yes, 0 = No}	
Stage 2 Model	After the first semester	754	***All features from stage 1, plus***	Nominal
			CS111ASGN {0 = N, 1 = E, 2 = D, 3 = C, 4 = C+, 5 = B, 6 = B+, 7 = A, 8 = A+}	Nominal
			CS111QUIZ {0 = N, 1 = E, 2 = D, 3 = C, 4 = C+, 5 = B, 6 = B+, 7 = A, 8 = A+}	Nominal
			CS111FPOST{0 = None, 1 = Low, 2 = Moderate, 3 = High}	Nominal
			CS111CPACC {0 = None, 1 = Low, 2 = Moderate, 3 = High}	Nominal
			CS111CW {0 = N, 1 = E, 2 = D, 3 = C, 4 = C+, 5 = B, 6 = B+, 7 = A, 8 = A+}	Nominal
			CS111GRADE {0 = EX, 1 = E, 2 = D, 3 = C, 4 = C+, 5 = B, 6 = B+, 7 = A, 8 = A+}	Numeric
			S1UNITS {1–4}	Nominal
			DROPOUT {1 = Yes, 0 = No}	
Stage 3 Model	After the second semester	502	***All features from stage 1 and stage 2, plus***	Nominal
			CS112ASGN {0 = N, 1 = E, 2 = D, 3 = C, 4 = C+, 5 = B, 6 = B+, 7 = A, 8 = A+}	Nominal
			CS112CPACC {0 = None, 1 = Low, 2 = Moderate, 3 = High}	Nominal
			CS112CW {0 = N, 1 = E, 2 = D, 3 = C, 4 = C+, 5 = B, 6 = B+, 7 = A, 8 = A+}	Nominal
			CS112GRADE {0 = EX, 1 = E, 2 = D, 3 = C, 4 = C+, 5 = B, 6 = B+, 7 = A, 8 = A+}	Nominal
			DROPOUT {1 = Yes, 0 = No}	

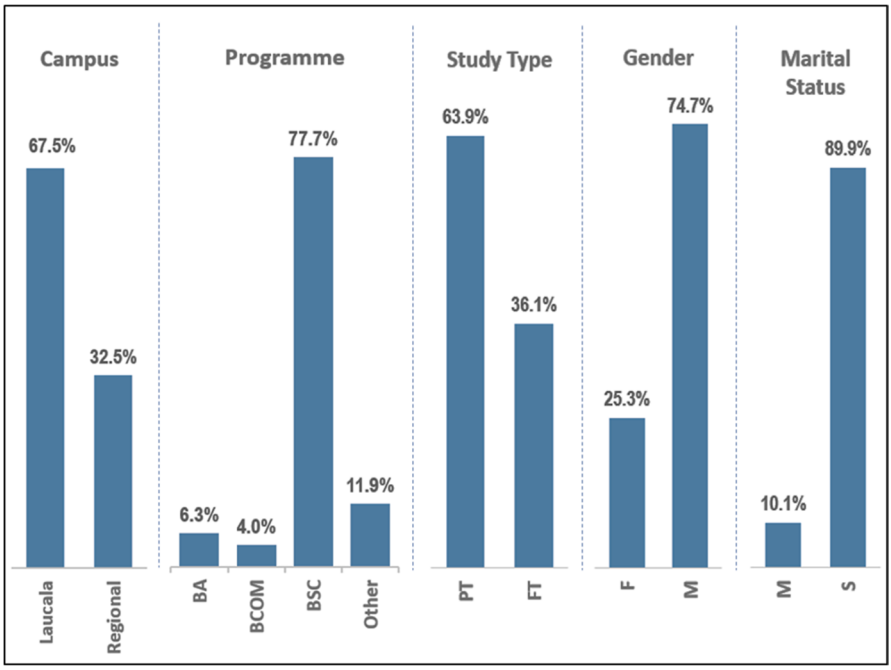

Fig. 2 The demographic profile of the students selected in the sample

to as the minority class). The data mining algorithms are built on the assumption of equal distribution of the classes, but in an imbalanced dataset, the algorithms become biased towards the majority class. However, most of the classification tasks require highly accurate prediction of the minority class, thus minimizing the misclassification of the minority class is critical. The high accuracy, in this case, will be misleading as this will not be reflective of the model's ability to minimize the error of classifying the minority classes incorrectly. The SMOTE, proposed by Chawla et al. in (2002), is applied to avoid the problem of overfitting in this study. This technique does not create duplicates; instead, it selects a subset from the minority class that are used as examples for creating new synthetic similar instances. The new synthetically created instances are then appended in the minority class to achieve balance. The current study utilized the SMOTE technique because it has been found to be an effective method for overcoming the class imbalance problem in the predictive modelling of student attrition (Márquez-Vera et al., 2013; Rovira et al., 2017; Thammasiri et al., 2014).

3.3 Modelling

The tool used for the data mining model creation is the open-source statistical software, R (version 3.4.4) which is widely used among statisticians and data miners. The following are some R packages and functions used in this project:

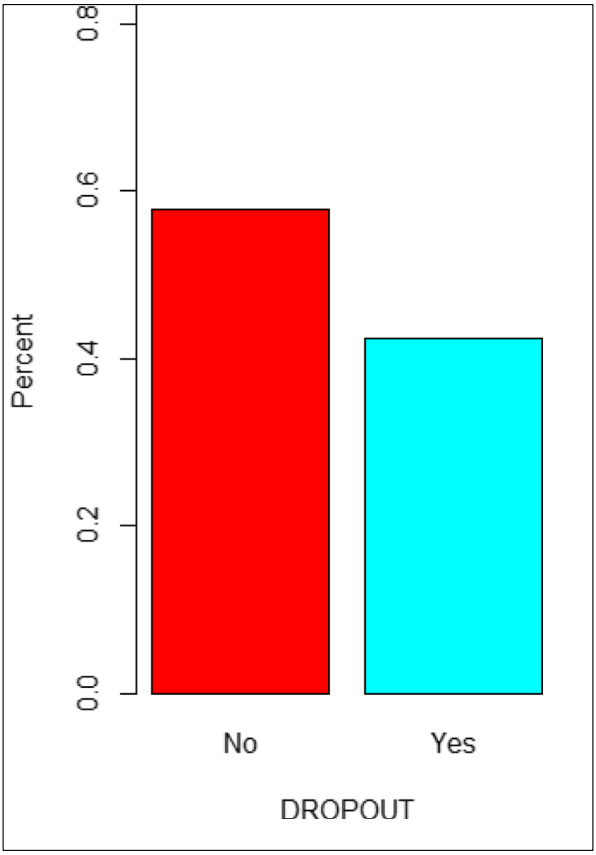

Fig. 3 Class distribution of the target variable for model 1

- ggplot2 – for data visualization
- Boruta – for feature selection
- Caret – for machine learning algorithm
- rpart – for building classification trees

The classification algorithms used in this research were the Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, and K-Nearest Neighbour. The selection was influenced by the findings of the literature survey, which suggested that they were more popular amongst researchers studying student attrition and that these algorithms had produced better performance accuracies while predicting student dropout. All the models were trained and tested using tenfold cross-validation with repetition of three.

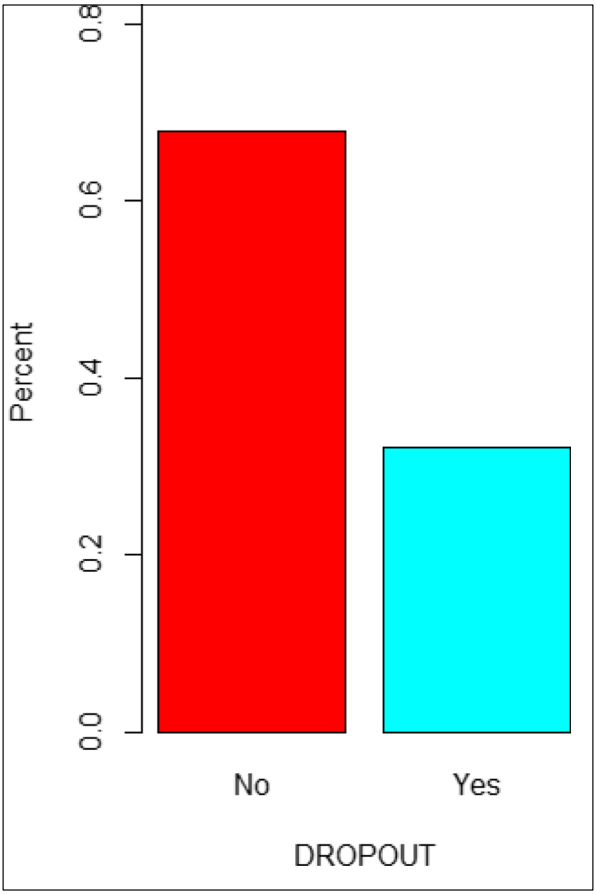

Fig. 4 Class distribution of the target variable for model 2

3.3.1 Decision Tree

Decision tree is a supervised learning technique that is suitable for classification as well as regression problems. The recursive partitioning and regression trees (**rpart**) method was used to train and test the decision tree model. The splitting index used for classification splitting was changed from the default which was *'gini'* to *'information'* to get the impurity in the class variables.

3.3.2 Random Forest

A random search was implemented using the **rf** method from the *Caret* function in R to build the Random Forest model. A tuning length of 15 was selected to identify the optimal parameters for *mtry*.

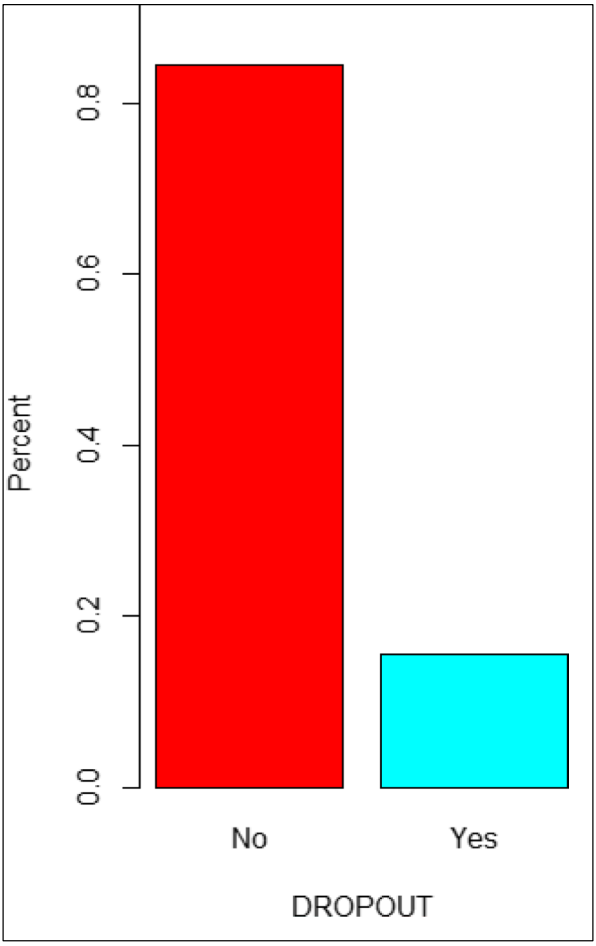

Fig. 5 Class distribution of the target variable for model 3

3.3.3 Naïve Bayes

Naïve Bayes is a probabilistic algorithm based on the Bayesian Theorem which operates on the conditional probabilities. It is used for a variety of classification tasks including the prediction of student attrition. No tuning of hyperparameters was applied to the Naïve Bayes model.

3.3.4 Logistic Regression

Logistic regression is a classification technique which is often used to model the probability of a dichotomous dependant variable, which was the case in this work. The generalized linear model (**glm**) of R was used in this study to build the

predictive model with *'binomial'* selected as the error distribution function. A tuning length of 5 was used to identify the optimal parameters.

3.3.5 K-Nearest Neighbour

The K-NN is another supervised learning algorithm which is used for both classification and regression tasks depending on the nature of the class variable. This technique uses distance metrics (for example, the Euclidean distance) to identify the closeness of an object to a class. An object which is identified to be in the proximity of a particular class is classified as that class. The grid search technique was to identify the optimal value of k in this task.

3.4 Evaluation

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 accuracy of a model is given by the proportion of true positive and true negative in all evaluated cases. The formula to calculate the accuracy is represented as.

$$\textbf{\textit{Accuracy}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Where:

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

The ***sensitivity*** of a model is its ability to determine the dropout cases correctly. Sensitivity is also known as the True Positive Rate (TPR) or recall. It represents the percentage of dropout cases that are correctly classified. The formula to calculate the sensitivity is represented as

$$\textbf{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

The ***specificity*** of a model is its ability to determine the non-dropout cases correctly. It is also known as the True Negative Rate (TNR), which is the percentage of non-dropout cases correctly classified as non-dropout. The formula to calculate the specificity is represented as

$$\textbf{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4}$$

Other common performance metrics include ***Receiver Operating Characteristic*** (ROC) and ***Area Under the Curve*** (AUC). The ROC curve is a method of comparing the different models holistically where the sensitivity is plotted against 1-specificity. The AUC is another measure of model performance which represents the models ability to distinguish between classes; the greater the AUC, the better the model's performance.

4 Results

The dataset for all the three stages were randomly partitioned into train and test sets using the 70%-30% split. The training data used for training the model consisted of 70% instances from the dataset while the remaining 30% was used for testing the model. Since the three datasets contained imbalanced target class, SMOTE balancing was applied to training data to ensure that the target class is equally distributed.

4.1 Feature selection

Feature selection was carried out using the Boruta algorithm in R (Kursa & Rudnicki, 2010) which is built on the basis of wrapper method (Chandrashekar & Sahin, 2014) where a subset of features are used to train the model, and then the inferences made from the previous model helps in deciding whether to add or drop feature(s) from the subset. The dataset for the first stage models contained 13 attributes and a target variable. Table 3 shows the results of the feature selection for model 1 using the Boruta algorithm. Three attributes were rejected by Boruta due to their low importance. Amongst the rejected attributes were INTC, PROG and DIS. The feature that was found to be the most important was STYPE having the highest mean and median importance scores. MGRADE was also one of the important attributes

Table 3 Feature selection for stage 1 model

	meanImp	medianImp	minImp	maxImp	normHits	decision
STYPE	**31.04621**	**30.85766**	**26.0962**	**35.98645**	**1**	**Confirmed**
MGRADE	**14.16634**	**13.86887**	**9.918376**	**19.61391**	**1**	**Confirmed**
CAMP	**8.366059**	**8.255576**	**3.232177**	**12.37273**	**1**	**Confirmed**
SPON	**7.84199**	**8.037389**	**2.686222**	**12.33361**	**0.981818**	**Confirmed**
NATION	**7.823497**	**7.788215**	**3.89491**	**12.03895**	**0.981818**	**Confirmed**
INCOUNT	**6.936169**	**6.889816**	**3.818073**	**10.38609**	**0.954546**	**Confirmed**
HSECLEV	**5.99441**	**5.995909**	**1.061105**	**9.71124**	**0.918182**	**Confirmed**
GEN	**5.704451**	**5.558055**	**0.847149**	**12.09616**	**0.818182**	**Confirmed**
AGE	**5.22287**	**5.24152**	**0.924493**	**9.093151**	**0.836364**	**Confirmed**
MSTAT	**4.124101**	**4.086351**	**1.625608**	**7.838725**	**0.654546**	**Confirmed**
INTC	2.036468	1.957493	-1.02088	5.85636	0.118182	Rejected
PROG	1.891669	1.756325	-0.35132	4.712962	0.118182	Rejected
DIS	-1.90684	-2.00738	-3.3528	-0.48097	0	Rejected

Table 4 Feature selection for stage 2 model

	meanImp	medianImp	minImp	maxImp	normHits	decision
CS111GRADE	**26.77422**	**26.78909**	**23.00772**	**29.82661**	**1**	**Confirmed**
CS111CW	**22.3812**	**22.27173**	**19.33814**	**25.54467**	**1**	**Confirmed**
CS111ASGN	**12.04399**	**12.07268**	**8.344685**	**16.26205**	**1**	**Confirmed**
STYPE	**10.59802**	**10.56931**	**7.86411**	**14.17705**	**1**	**Confirmed**
CS111CPACC	**10.07895**	**10.0067**	**6.349559**	**13.00435**	**1**	**Confirmed**
HSECLEV	**5.294147**	**5.298507**	**2.499181**	**8.434802**	**0.917197**	**Confirmed**
MGRADE	**4.533859**	**4.643046**	**1.158257**	**8.509037**	**0.843949**	**Confirmed**
MSTAT	**4.293596**	**4.319706**	**0.882993**	**7.114997**	**0.812102**	**Confirmed**
CAMP	**4.095524**	**4.135449**	**0.963863**	**7.48145**	**0.745223**	**Confirmed**
AGE	**3.619957**	**3.607935**	**-0.03877**	**6.88905**	**0.66879**	**Confirmed**
S1UNITS	**3.396227**	**3.39883**	**0.507296**	**6.476563**	**0.617834**	**Confirmed**
NATION	**3.284297**	**3.279819**	**-0.7673**	**6.310996**	**0.595541**	**Confirmed**
CS111QUIZ	**3.257322**	**3.212337**	**-0.59193**	**6.668128**	**0.61465**	**Confirmed**
INTC	1.843384	1.898228	-0.63144	4.053666	0.025478	Rejected
INCOUNT	1.643822	1.406886	0.124639	3.677485	0.012739	Rejected
SPON	1.325019	1.584412	-0.8962	3.253535	0.022293	Rejected
GEN	0.887301	0.864858	-1.80294	2.652987	0.006369	Rejected
CS111FPOST	0.376068	0.574008	-2.40725	1.475355	0	Rejected
PROG	-0.31912	-0.01145	-2.26142	1.142443	0	Rejected
DIS	-2.73045	-2.87811	-3.40706	-1.65747	0	Rejected

for stage 1 models, which shows that low Math grade achieved in high school can significantly influence students' decision to drop out from their Computing Science degrees. Hence, high school math grades must be considered by academic advisors for recommending students to do a Computing Science degree programme.

Models for stage 2 were built using the second dataset consisting of demographic, financial, prior educational background features and the first-semester academic variables related to CS111. The important features selected by the feature selection algorithm for model 2 are shown in Table 4. Variables relating to the academic performance in CS111 were found to be the most influential features. The grade received in CS111 was the most important which shows that there is a lower likelihood of a student returning to enrol in Computing Science in the second semester after failing their first programming course. Other academic performance-related variables that were identified as influential in the second model are assessment marks such as assignment and quiz marks obtained by the student. The number of courses taken by the student in the first semester of study and the frequency of access to Moodle course page were also included in the list of important predictors. Important demographic variables included STYPE, MSTAT, CAMP, AGE and NATION. Both prior education attributes were also found to be important predictors of student drop out in model 2.

Table 5 Feature selection for stage 3 models

	meanImp	medianImp	minImp	maxImp	normHits	decision
CS112GRADE	**23.24552**	**23.42458**	**16.05154**	**27.7152**	**1**	**Confirmed**
HSECLEV	**8.57654**	**8.573295**	**4.963813**	**11.89471**	**0.995992**	**Confirmed**
CS111QUIZ	**4.8822**	**4.856658**	**1.129891**	**8.54127**	**0.87976**	**Confirmed**
AGE	**4.821684**	**4.823265**	**1.295562**	**8.105087**	**0.893788**	**Confirmed**
CS111GRADE	**3.924159**	**3.964373**	**-0.07306**	**6.452096**	**0.791583**	**Confirmed**
MGRADE	**3.908035**	**3.947065**	**-0.47208**	**7.626248**	**0.759519**	**Confirmed**
CS112ASSGN	**3.690464**	**3.741122**	**0.25839**	**7.007281**	**0.737475**	**Confirmed**
STYPE	**3.594207**	**3.581447**	**-0.87843**	**7.235642**	**0.715431**	**Confirmed**
CS111CW	**3.278485**	**3.316074**	**-0.47481**	**7.18988**	**0.647295**	**Confirmed**
NATION	**2.427526**	**2.435475**	**-1.53353**	**5.90762**	**0.466934**	**Tentative**
S1UNITS	1.966471	2.012645	-1.52088	4.658222	0.152305	Rejected
CS111ASGN	1.894554	1.895472	-1.47411	5.113383	0.084168	Rejected
CS112CW	1.842609	1.969823	-0.38985	3.985084	0.064128	Rejected
CAMP	1.735627	1.666259	-2.04212	4.547268	0.076152	Rejected
INTC	1.671075	1.806239	-0.87589	3.345633	0.04008	Rejected
CS111FPOST	1.130659	0.965497	-0.17016	2.760634	0.004008	Rejected
INCOUNT	0.912979	0.964098	-0.51401	2.487419	0.002004	Rejected
CS112CPACC	0.803665	0.573042	-0.98285	2.985167	0.006012	Rejected
GEN	0.657432	0.363621	-0.96161	2.50284	0.004008	Rejected
MSTAT	0.433108	0.58934	-2.28264	3.054878	0.002004	Rejected
SPON	0.341136	0.21409	-0.80892	1.410409	0	Rejected
PROG	0.129777	-0.26801	-1.24239	1.919431	0	Rejected
DIS	0	0	0	0	0	Rejected
CS111CPACC	0	0	0	0	0	Rejected

Using 500 iterations, the Boruta algorithm identified 10 attributes as important for the stage 3 models. Table 5 shows the attribute importance list produced by the feature selection algorithm.

4.2 Performance evaluation

Figure 6 shows the accuracy of the five classifiers for the three predictive models. The graph shows that the third model (model 3) had the best accuracy between all classifiers except the DT model for which the second model had a higher accuracy. Model 1 had the lowest accuracy for all classifiers ranging between 50–60%. The LR classifier in model 3 achieved the best predictive accuracy classifying 80% of the observations correctly.

As discussed earlier, that accuracy can be misleading, especially when a large number of negative cases are correctly classified compared to the positive cases (that is, when TNR is high, and TPR is relatively low). This

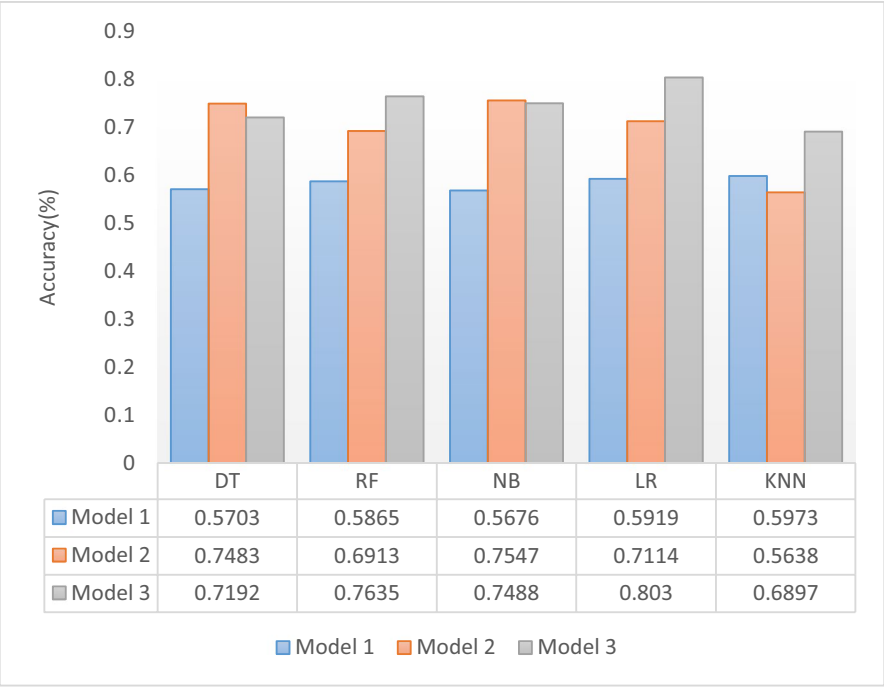

Fig. 6 Accuracy of the three models

Table 6 Precision scores for the three models

	Model 1	Model 2	Model 3
Decision Tree	44%	53%	28%
Random Forest	49%	44%	26%
Naïve Bayes	47%	46%	31%
Logistic Regression	50%	47%	40%
K-Nearest Neighbour	50%	35%	23%

Table 7 Recall scores for the three models

	Model 1	Model 2	Model 3
Decision Tree	21%	51%	68%
Random Forest	67%	66%	39%
Naïve Bayes	48%	68%	64%
Logistic Regression	43%	68%	86%
K-Nearest Neighbour	64%	72%	54%

Table 8 F1 scores for the three models

	Model 1	Model 2	Model 3
Decision Tree	29%	51%	40%
Random Forest	57%	53%	31%
Naïve Bayes	47%	55%	41%
Logistic Regression	46%	56%	55%
K-Nearest Neighbour	56%	47%	32%

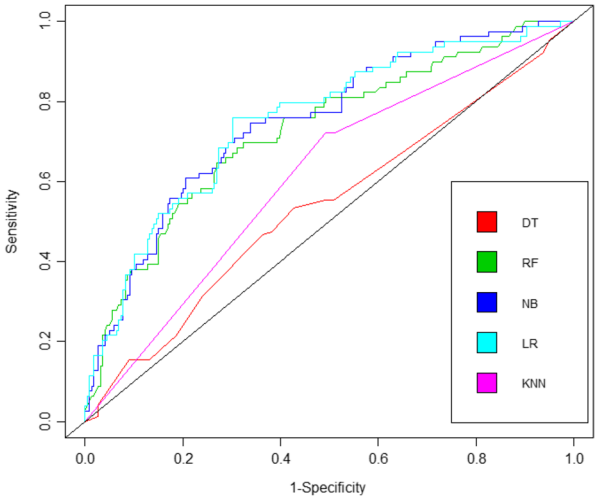

Fig. 7 ROC curves for model 1

can become problematic if the cost of false negative is high. For this reason, other metrics such as precision, recall, F1, ROC and AUC must be considered while evaluating the model performance to determine the best model.

The precision scores of the five classifiers for the three models are presented in Table 6. Precision score indicates how well a model is able to classify a negative class (non-dropout cases) while recall (sensitivity) is the measure of how well a model correctly identifies the positive class (dropout cases). When comparing the precision scores of the five classifiers, it can be noted that the DT classifier in model 2 achieved the highest score with 53% negative instances correctly classified.

Since the aim of this study is to achieve better predictive accuracy for the positive class (dropout cases), it is critical to compare the recall scores of the five classifiers for the three models which are provided in Table 7. The Random Forest and K- Nearest Neighbour (KNN) classifiers had a high recall score for the models in the first stage. The RF classifier performed the best at predicting the dropout cases with 67% accuracy. The KNN classifier for model 2 had predicted 72% drop out cases correctly, which was the best performance in stage 2. LR classifier

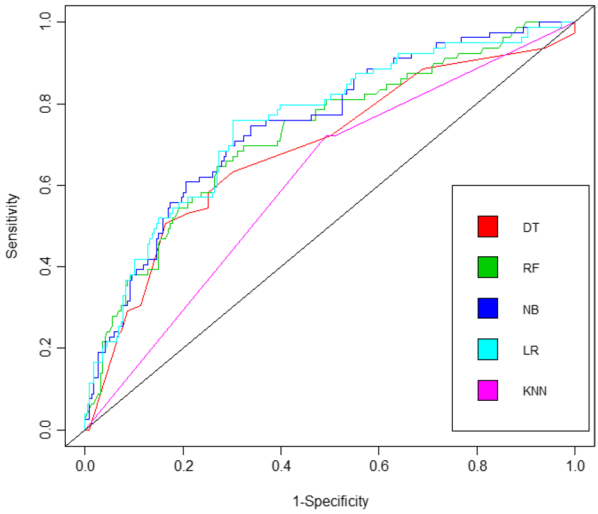

Fig. 8 ROC curves for model 2

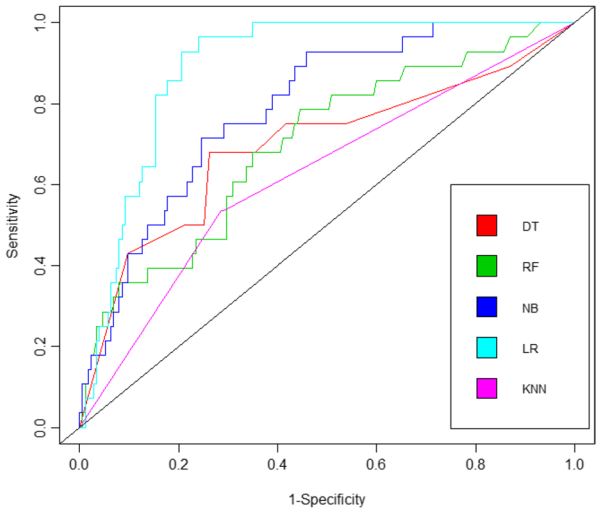

Fig. 9 ROC curves for model 3

in model 3 had achieved high recall score of 86% outperforming all other classifiers in stage 3 in terms of correctly classifying the positive class.

The F1 scores of the five classifiers for the three models are shown in Table 8. The RF classifier in model 1 achieved the highest F1 score of 57%, but the difference between some other classifiers such as the KNN classifier in model 1 and the LR classifier in model 2 and model 3 was not large.

With the results of the accuracy, precision, recall and F1 scores, it is still not clear as to which of the five classifiers produced the best model and there is not enough evidence to select the best time to run the predictive model. Therefore, the ROC curves and AUC scores must be compared to determine the best model

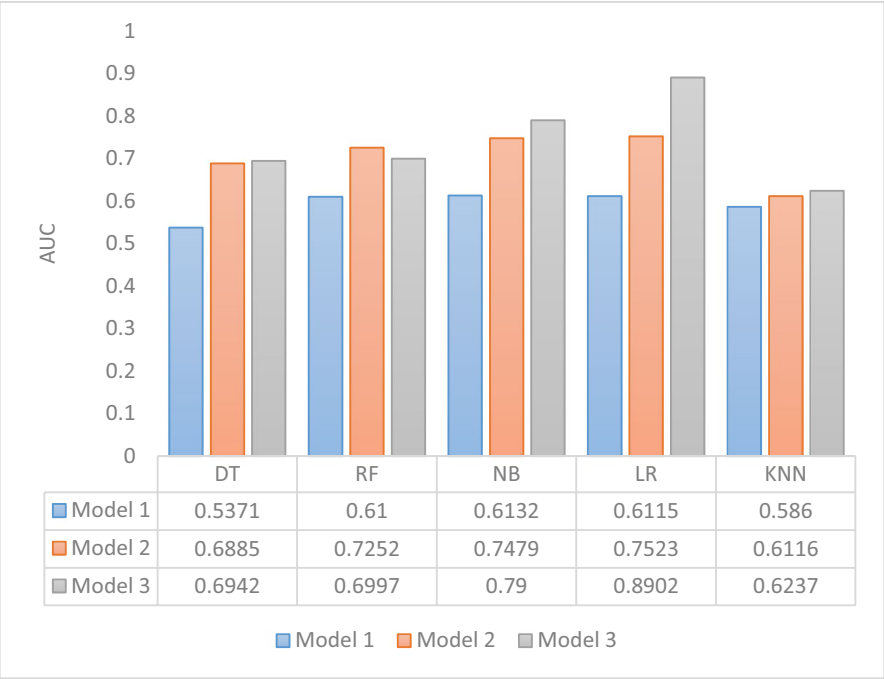

Fig. 10 Area under the ROC curve for the three models

and the most suitable time to run the model. To further test the performance of the models, ROC curves and AUC were used. Figure 7, 8, and 9 present the ROC curves of the five classifiers in models 1, 2 and 3, respectively. The ROC curves for the KNN and DT classifiers are nearer to the threshold line for all the three models. This means that these classifiers did not perform well compared to the other three classifiers. The ROC curves for RF, NB and LR classifiers in models 1 and 2 are very similar while in model 3 the ROC curve for the LR classifier is the furthest away from the threshold line as well as from the other classifiers.

AUC was used as another measure to identify the difference in the performance of the five classification algorithms in the three models. AUC quantifies the performance difference between classifiers. Figure 10 presents the AUC scores for the five classifiers.

In stage one, the NB classifier achieved the best performance, but the RF and LR also displayed comparable results. The AUC scores are 0.6132, 0.6100 and 0.6115, respectively. For the second stage, the RF, NB and LR again achieved high AUC scores, but the LR model was able to achieve the best predictive performance with an AUC score of 0.7523. The LR classifier had an AUC score of 0.8902, which was the best performance in the third stage. Additionally, it can be noted that the LR classifier in stage 3 also had better accuracy and recall scores compared to the other classifiers. The better performance of the models in stage 3 is due to the availability of more data during this stage.

5 Discussion

The findings are in line with previous work related to predictive modelling of student attrition. For example, Delen (2012) achieved accuracies ranging between 74 and 81% with the best performance displayed by the ANN classier model. In another study, Aulck et al. (2017) found that the LR model produced the best accuracy of 81% which is comparable to the results of this research with 80.3% accuracy which was also achieved by the LR classifier model in stage 3. The models in this research achieved better prediction accuracy compared to the models built by Kovacic (2010) which obtained highest accuracy of 61%. The results show that data mining techniques are capable of producing acceptable accuracies when predicting student attrition at discipline level.

The confirmed attributes are the ones that have a higher contribution in the decision. It can be depicted from Table 3 that the attributes corresponding to the students' performance in the two first-year courses in CS programmes are the ones that are given the highest importance. Additionally, it can be noted that students who perform poorly in CS112 have a higher likelihood of dropping out of their programmes. These results are supported by studies conducted by (Delen, 2012; Giannakos et al., 2017; Orozco & Niguidula, 2017) which reveal that students who get good grades in early courses such as those taken in the first two semesters are more likely to return and complete their programmes within the prescribed time. Unsatisfactory academic performance and poor grades in early courses can influence the decision to drop out in three ways:

1. First-year courses are usually prerequisites for most of the CS courses in the sophomore year. Therefore, students who get a fail grade in these courses are not able to progress to the next level.
2. Students who persistently get a failing grade in these courses usually end up with a lower Grade Point Average (GPA), which can also affect the students' progression.
3. Failing a grade can be costly, and if a student makes multiple attempts then this will incur more cost. Often these costs are borne by parents of students who are studying privately so those who are not able to contend the additional expenses are more likely to drop out. Some sponsors also allow students to repeat a course only a specific number of times. Students who fail these courses more than the specified number of times are not able to afford the fees so they may also drop out.

Other academic performance related attributes that were identified as influential are assessment marks such as assignment and quiz marks obtained by the student and the workload of the student in the first year of study.

With the increase in the use of LMS in HE majority of the learning and assessment materials are made available online. Through the use of LMS such as Moodle, students are able to access the course pages to view the topic lessons, attempt online quizzes, participate in discussion forums and submit assignments in online dropboxes. Students who do not regularly interact with Moodle have higher chance of

poor performance in the courses. Thus, the academic success of students are hugely influenced by their online presence in the respective courses. The attributes that relate to the students online presence in this study include the number of assignments submitted, number of quizzes attempted, total number of posts in Moodle discussion forums and the frequency of access of Moodle course page.

Demographic attributes such as age and study type were also found to be important while gender, marital status, programme, campus and disability were not. Students who transfer directly from high school or foundation studies into HE are often more committed to studies. These students usually get more time to study as they spend more time on campus. On the contrary, many students who return to study after a lapse often face difficulties in coping with the courses since these students have work and family commitments. Due to these reasons most of the older students often fail to perform well, consequently withdrawing from their studies.

The attributes corresponding to a student's previous academic performance such as high school math grade and highest level of secondary school attended were also amongst the important attributes. A good math background is fundamental in programming courses (Beaubouef, 2002). Students who lack problem solving skills often struggle to succeed in understanding algorithms. This explains the inclusion of high school math grade in the subset of influential factors.

The use of machine learning algorithms such as the ones implemented in this work can enable the detection of at-risk students with high accuracies. The logistic regression models have been noted to perform well with the data of mixed variable type. Due to the imabalanced nature of the dropout attribute, it is highly recommended to utilize a resampling technique such as SMOTE. The performance measures such as accuracy are more reliable with balanced datasets. Furthermore, the use of predictive modelling at different stages helps in understanding the different contributors of student drop out. Having the knowledge of which student is more likely to drop out of a programme enables the academic staff as well as other relevant stakeholders to plan proper intervention strategies in a more efficient manner.

6 Conclusion

This study used data mining techniques for predictive modelling of freshmen student attrition in Computing Science at a regional university in the South Pacific. Having apriori knowledge of the set of students who would drop out from a programme can help decision-makers to concentrate their attention to this cohort and provide them with proper guidance and intervention programmes to retain them. Three models at three different stages of the first-year studies were built because there was a need to find what features are more influential at the different stages of learning. For this reason, data were divided into three subsets; one for each stage with different features involved in each subset. The student attrition data is usually imbalanced as the number of students dropping out is fewer compared to those who are retained. Data resampling using SMOTE technique was carried out to improve the misclassification of the false negatives (dropout cases identified as non-dropout), as the cost of a greater false negative rate is higher. For the non-dropout students who are

incorrectly classified as dropouts will also be benefitting from the intervention programmes, further improving the chance of their academic success. Feature selection was applied to the three datasets to find the most important predictors.

Five classification algorithms were used to train and test the three models. These consisted of the Decision Tree, Random Forest, Naive Bayes, Logistic Regression and k-Nearest Neighbour. Several measures were used to compare the performance of these models but the most reliable metric was found to be the ROC and AUCROC which gave a true picture of how the models performed at predicting the positive and the negative classes correctly. In stage 1, the NB model had best performance while the LR models in stage 2 and 3 outperformed all its couterparts with an AUC of 75% and 89%, respectively.

Acknowledgements The partial data in the result was presented at the 2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) in Melbourne, VIC, Australia.

Anonymised raw data used in this study can be obtained from this link: https://github.com/mohd-naseem/Student-Attrition/blob/main/Student%20attrition.csv

Declarations

Conflict of Interest None.

References

Aguiar, E., Chawla, N. V., Brockman, J., Ambrose, G. A., & Goodrich, V. (2014). Engagement vs performance: using electronic portfolios to predict first semester engineering student retention. *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (pp. 103–112). ACM.

Al-Badarenah, A., & Alsakran, J. (2016). An automated recommender system for course selection. *International Journal of Advanced Computer Science and Applications, 7*(3), 1166–1175.

Aulck, L., Aras, R., Li, L., L'Heureux, C., Lu, P., & West, J. (2017). STEM-ming the Tide: Predicting STEM attrition using student transcript data. *arXiv preprint arXiv:1708.09344*.

Badr, G., Algobail, A., Almutairi, H., & Almutery, M. (2016). Predicting students' performance in university courses: a case study and tool in KSU mathematics department. *Procedia Computer Science 82*, (pp. 80–89).

Baker, R. S., & Kalina, Y. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM Journal of Educational Data Mining, 1*(1), 3–17.

Beaubouef, T. (2002). Why computer science students need math. *SIGCSE Bulletin, 34*(4), 57–59.

Blekic, M., Carpenter, R., & Cao, Y. (2017). Continuing and transfer students: Exploring retention and second-year success. *Journal of College Student Retention: Research, Theory & Practice, 22*(1), 71–98. https://doi.org/10.1177/1521025117726048

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16–28.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Costa, E. B., Fonseca, B., Santana, M. A., de Araujo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247–256.

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *International Working Group on Educational Data Mining*. Cordoba, Spain.

Delen, D. (2012). Predicting Student Attrition with Data Mining Methods. *Journal of College Student Retention: Research, Theory & Practice, 13*(1), 17–35.

Dolatabadi, S. H., & Keynia, F. (2017). Designing of customer and employee churn prediction model based on data mining method and neural predictor. *In 2017 2nd International Conference on Computer and Communication Systems (ICCCS)* (pp. 74–77). IEEE.

Evans, M. (2000). Planning for the transition to tertiary study: A literature. *Journal of Institutional Research, 9*(1), 1–13.

Gairín, S. J., i Ivern, T., Ma, X., Feixas Condom, M., Gazo, P., Aparicio Chueca, M., & Torrado Fonseca, M. (2014).Student dropout rates in Catalan universities: Profile and motives for disengagement. *Quality in Higher Education* 20(2), 165-182

Ghadeer, A.-O.S., & Alaa, E.-H.M. (2015). Data Mining In Higher Education: University Student Dropout Case Study. *International Journal of Data Mining & Knowledge Management Process (IJDKP), 5*(1), 15–27.

Giannakos, M. N., Pappas, I. O., Jaccheri, L., & Sampson, D. G. (2017). Understanding student retention in computer science education: The role of environment, gains, barriers and usefulness. *Education and Information Technologies, 22*(5), 2365–2382.

Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer Segmentation using K-means Clustering. *In 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS). 7*, pp. 135–139. IEEE.

Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the changing trends of market data using association rule mining. *Procedia Computer Science, 85*, 78–85.

Kazemi, A., Babaei, M. E., & Javad, M. O. (2015). A data mining approach for turning potential customers into real ones in basket purchase analysis. *International Journal of Business Information Systems, 19*(2), 139–158.

Kemper, L., Vorhoff, G., & Wigger, B. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education, 10*(1), 28–47.

Kori, K., Margus, P., Eno, T., Tauno, P., Heilo, A., Ramon, R., . . . Tiia, R. (2015). First-year dropout in ICT studies. *2015 IEEE Global Engineering Education Conference (EDUCON)* (pp. 437–445). IEEE.

Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data. *Informing Science + Information Technology Education Joint Conference*. Cassino, Italy. Retrieved from http://hdl.handle.net/11072/646

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software, 36*(11), 1–13.

Lacave, C., Molina, A. I., & Cruz-Lemus, J. A. (2018). Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks. *Behaviour & Information Technology*, (pp. 1–15).

Lin, C. F., Yeh, Y. C., Hung, Y. H., & Chang, R. I. (2013). Data mining for providing a personalized learning path in creativity: An application of decision trees. *Computers & Education, 68*, 199–210.

Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence, 38*(3), 315–330.

Minges, M., & Stork, C. (2015). *Economic and social impact of ICT in the Pacific*. Pacific Region Infrastructure Facility.

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education, 40*(3), 355–371.

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., & Verbeke, W. (2020). Uplift Modeling for preventing student dropout in higher education. *Decision Support Systems, 134*, 113320.

Orozco, M. E., & Niguidula, J. C. (2017). Predicting Student Attrition Using Data Mining Predictive Models. *Proceedings of 143rd The IIER International Conference.* Jeju Island, South Korea.

Oztekin, A. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors. *Industrial Management & Data Systems, 116*(8), 1678–1699.

Pal, S. (2012). Mining Educational Data to Reduce Dropout Rates of Engineering Students. *International Journal of Information Engineering and Electronic Business, 2*, 1–7.

Patil, R., Salunke, S., Kalbhor, M., & Lomte, R. (2018). Prediction System for Student Performance Using Data Mining Classification. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1–4). IEEE.

Pérez, B., Castellanos, C., & Correal, D. (2018). Predicting student drop-out rates using data mining techniques: A case study. *IEEE Colombian Conference on Applications in Computational Intelligence* (pp. 111–125). Springer, Cham.

Reddy, P., & Sharma, B. (2015). Effectiveness of Tablet Learning in Online Courses at University of the South Pacific. *Proceedings of Asia-Pacific World Congress on Computer Science and Engineering* (pp. 1–9). Fiji: IEEE.

Reddy, E., & Sharma, B. (2018). Mobile Learning Perception and Attitude of Secondary School Students in the Pacific Islands. *Proceedings of the 22nd Pacific Asia Conference on Information Systems (PACIS 2018).* Yokohama, Japan. Retrieved from https://aisel.aisnet.org/pacis2018/319/

Richards, E., & Terkanian, D. (2013). Occupational employment projections to 2022. *Monthly Labor Review, 136*, 1.

Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PLoS ONE, 12*(2), e0171207. https://doi.org/10.1371/journal.pone.0171207

Schneider, K., Berens, J., Oster, S., & Burghoff, J. (2018). Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods. *Annual Conference 2018 (Freiburg, Breisgau): Digital Economy.* Verein für Socialpolitik / German Economic Association. Retrieved from https://ideas.repec.org/p/zbw/vfsc18/181544.html

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

Sharma, B., Kumar, R., Rao, V., Finiasi, R., Chand, S., Singh, V., & Naicker, R. (2017). A Mobile Learning Journey in Pacific Education. In *Angela Murphy et al. (Eds) Mobile Learning in Higher Education in the Asia-Pacific Region – Harnessing Trends and Challenging Orthodoxies* (Vol. 40, pp. 581–606).

Shilbayeh, S., & Abonamah, A. (2021). Predicting Student Enrolments and Attrition Patterns in Higher Educational Institutions using Machine Learning. *International Arab Journal of Information Technology, 18*(4), 562–567.

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange, 1*(1), 64–85.

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications, 41*(2), 321–330.

Tinto, V. (1975). Dropout from Higher Education: A theatrical synthesis of recent research. *Review of Education Research, 45*, 89–125.

Uliyan, D., Aljaloud, A. S., Alkhalil, A., Al Amer, H. S., Mohamed, M. A., & Alogali, A. F. (2021). Deep Learning Model to Predict Students Retention Using BLSTM and CRF. *IEEE Access, 9*, 135550–135558.

Yaacob, W. W., Sobri, M., Nasir, S. M., Norshahidi, N. D., & Husin, W. W. (2020). Predicting student drop-out in higher institution using data mining techniques. *Journal of Physics: Conference Series, 1496*(1), 1–13.

Yasmin, D. (2013). Application of the classification tree model in predicting learner dropout behaviour in open and distance learning. *Distance Education, 34*(2), 218–231.

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year. *Journal of Data Science, 8*, 307–325.

Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and e-Learning, 17*(1), 118–133.

Zaffar, M., Hashmani, M. A., & Savita, K. S. (2018). A Study of Prediction Models for Students Enrolled in Programming Subjects. *2018 4th International Conference on Computer and Information Sciences (ICCOINS)* (pp. 1–5). IEEE.

Authors and Affiliations

Mohammed Naseem[1] · **Kaylash Chaudhary**[1] · **Bibhya Sharma**[1]

Kaylash Chaudhary
kaylash.chaudhary@usp.ac.fj

Bibhya Sharma
bibhya.sharma@usp.ac.fj

[1] School of Information Technology, Engineering, Mathematics and Physics, The University of the South Pacific, Suva, Fiji