Using Machine Learning to predict the completion of a learner's Undergraduate Science Degree based on their First-year Marks

Prince Ngema
Supervised by DR Ritesh Adjhooda
BSC Hons Big Data Analytics - 2019
School of Computer Science and Applied Mathematics

***Abstract*— Advances in the field of Machine learning have led to major improvements to Higher educational institution's attempts to predict student performance. The capacity to predict the performance of students is of utmost significance to any Higher Education Institution aiming to improve the performance of students and increase graduation rates. Students at risk of academic retention can be identified and be provided with the necessary support timeously. Machine Learning can be used in the development of predictive models that can be used to predict students' performance. However, predicting students' performance is a taxing task because student's academic performance is dependant on a number of factors. How a student performs in their first year of study may give a vivid idea of where they are headed academically. We can therefore use machine learning techniques to leverage these marks and predict students' performance. In this study, taking a conceptual framework proposed by Spady [1970] as a rationale, we attempt to predict (using first year marks) the completion of a learner's undergraduate Computer Science degree. For this cause, we employ several Machine Learning (ML) predictive models such as the J48, Naive Bayes, Support vector machines, Multilayer Perceptron, Logistic Regression and the k*. We focus on identifying the best performing ML predictive model for the task of predicting the completion of a learner's Computer science degree. We also focus on ranking the predictive power of Computer science courses taken by a learner in their first year of study at a South African institution. We also provide a software program that has the ability to predict the completion of a learner's Computer Science degree. With the 10-fold cross validation method, we compared the performance of the predictive models using their accuracy, precision, and recall values. The J48 predictive model achieved the highest accuracy value 0f 70% and for both precision and recall, the J48 predictive model achieved a value of 0.69. Comparison was also done using AUC. The J48 algorithm achieved the highest AUC value of 0.75. Using gain ratio, this study also revealed that a students who perform well in Calculus and Algebra is more likely to complete their studies in Computer Science.**

1. INTRODUCTION

Many a time, admittance into a University programme is a transformative experience for students, it gives them and their families hope for a lustrous future. Alas, some students who are accepted into the university programme fail to complete their degree due poor academic performances. These students are left lamenting and drowning in debt accumulated during their years of study. The early discovery of students who are at risk of not completing their studies out is of utmost importance to High education institutions[Osmanbegovic and Suljic 2012]. Early discovery of students at risk may result in increased students' performance, reduced attrition rates and increased retention rates. These changes ultimately lead to higher graduation rates [Aljohani 2016]. Fig 1 depicts goals of predicting students' performance. Many researchers

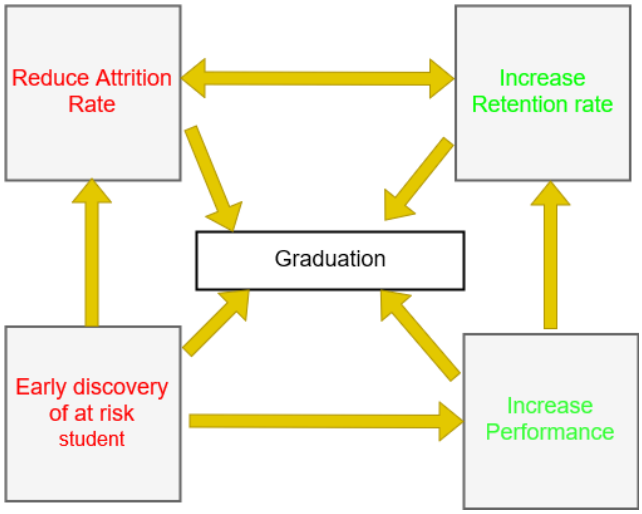


Fig. 1: goals of predicting student performance according to Aljohani [2016]

from different walks of life have developed models to predict students' academic performance. Studies show that the biggest attrition rate occurs at fresh-man level. In South Africa, 29% of students drop out after doing their fresh-man year studies and just 30% of fresh-man students graduate after five years [Scott et al. 2007]. This study is based on Computer Science learners at a South African Higher-Education Research-Intensive Institution between the years 2008 and 2017. These learners are categorized into three groups: "Completed", this where a learner completes their degree in minimum time (3 years); "Delayed", this when the learner completes their degree in more than 3 years; "Failed", this where the learner drops out and does not complete their degree. At this institution, they have been 21% of learners categorised as Completed; 24% as Delayed; and 55% as Failed. In light of these alarming figures, programs that will improve student retention rates and graduation rates are needed. These programs improve the performance of students by looking at students' previous performance and

using this knowledge to fortell how a student is likely to perform in their field of study [Yadav et al. 2012].

Previous work in field of predicting student performance has seen researchers employ different techniques like machine learning, statistical analysis and data mining in attempts of building models to do the prediction. In this study using Machine learning predictive models, we adopt a conceptual framework proposed by Spady [1970] as a rationale to predicting the completion of learner's Computer Science degree based on their first-year marks. According to Spady [1970], a student will decide to stay and complete their degree or dropout based on four factors: Grades, Intellectual development, normative congruence and friendship support. He further grouped these factors in to two systems, the Academic system and the Social system (see Fig 2). Grades and Intellectual development fall under the Academic system whereas normative congruence and friendship support fall under the social system. This study lies squarely within the academic system. Grades obtained during a student's first-year of study are taken as the predictor variable.

Following the conceptual framework of Spady [1970], we define several features associated with first-year marks that will be used to classify the student into three completion profiles: "Completed", "Delayed", and "Failed". During their first year of study, a student must have three majors and one elective module. These four constitute our feature space along with the student's degree completion outcome and the aggregate of all the marks obtain.

Different machine learning predictive models were used to predict the completion of the learner's degree. These classifiers are : J48, K*, Support Vector Machines, Logistic Regression, Naive Bayes, and Multilayer Perceptron. The goal of this study is exploring the possibility of applying these machine learning algorithms to predict the completion of a student's Computer Science degree based on first-year marks and investigating the value of using first-year marks as predictor variables. The research question can be seen as two fold:

- Are the above mentioned algorithms capable of predicting students' performance?
- Is using first-year marks as a predictor variables in a student performance prediction task viable?

Using 10-fold cross validation method, accuracy, precision, recall, AUC values and time taken to execute were used to scale the performance of these models. Information gain was used to rank features according to their predictive power. Based on the experiments it is found that the accuracy level of the classifiers range between 52% and 70%. The J48 achieved the biggest recorded accuracy. The AUC values vary between 0,7 to 0,8. The J48 algorithm had the highest AUC value. The J48 algorithm also took the least time to execute. In terms of Information gain, MajorOne is the highest ranked feature, that is, Calculus 1 and Algebra 1 have the highest deterministic power. An application software which uses the J48 predictive model to predict the completion of a learner's Computer Science degree has been prepared. The stake holders of this application are:

- Computer Science Students
- Companies that offer financial aid to students
- High learning Institution's administrators and teachers.

There are three major general contributions of this paper: (a) an interactive program which is able to predict the completion of a student's Computer Science degree (b) a comparison of various classification models to classify learner instances into these four completion outcomes; and (c) the first trained classifier able to calculate the probability of a learner completing their degree at a South African University focused on the the conceptual framework of Spady [1970].

This document is structured as follows. The Related Work highlights the contributions in the domain of predicting at-risk student profiles and a selected conceptual framework for student attrition; the Methodology highlights our data, feature selection, and choice of classification models; the Results outlines our major findings; and the Discussion and Conclusion summaries this paper, outlines our contributions, and puts forward recommendations of future work.

2. Related work

Numerous studies on students' performance have been made over the years. The goal is mostly to reduce attrition and increase graduation rates. Researchers employ different techniques such as data mining , statistical analysis and machine learning in attempts to predict students performance. This section presents a review of studies previously conducted in field of "student's performance prediction" related to the present study.

Many theoretical models (conceptual frameworks) on student performance have been developed to date. In this study we will adopt a conceptional framework proposed by Spady [1970] as our rationale. To explain the high dropout rates, Spady [1970] investigated the calibre of reciprocity between a student and their school's environment. He postulated that grades, friendship support, intellectual development and normative congruence affect a student's decision to stay and complete their degree or withdraw from an academic institution. He grouped these four factors into two systems (Academic system and Social system) as seen Fig 2. Grades and Intellectual development fall under the Academic system whereas normative congruence and friendship support fall under the social system. These four constitute input factors of the conceptual framework put forward by Spady [1970]. This study lies squarely within the academic system. We consider grades as the only input factor. There are many studies that have been conducted on predicting a learner's performance using grades obtained by the learner.

In this paper, predictive machine learning classifiers such as Naive Bayes, Logistic Regression, SVM, Multilayer Perceptron, J48 decision tree and K* are used to predict the performance of students based on their first year marks/grades. Student performance in this case implies the completion of a learner's undergraduate computer science

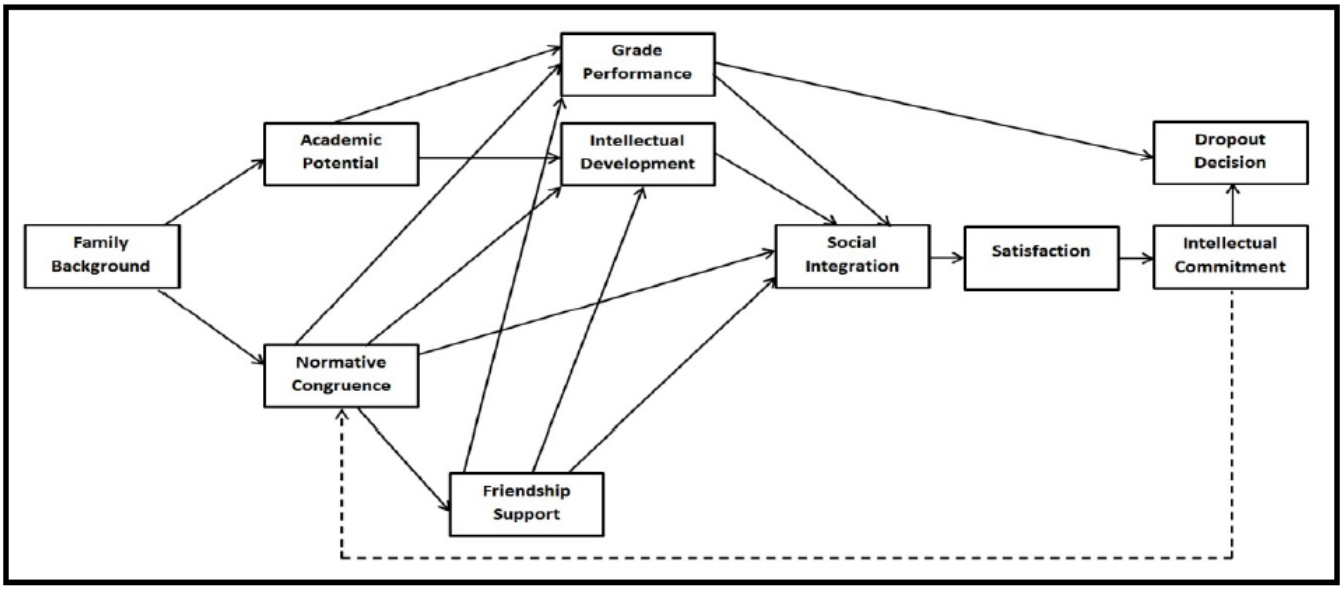


Fig. 2: Conceptual Framework proposed by Spady [1970]

degree. In Table I, studies related to our study are summarized systematically.

The use of grades as a factor that affects students' performance has been investigated by many researchers. In all of the studies reviewed in Table I, grades constitute the predictor variables. Studies by Abu-Oda and El-Halees [2015], Soule [2017], Sangodiah et al. [2015],Awad and Ewais [2018],Rusli et al. [2008] and Dekker et al. [2009] have grades as the sole predictor variable, these studies are more related to the present study.

In terms of the machine predictive algorithms used, the Naive Bayes was employed by Pojon [2017], Ajoodha and Jadhav [2019], Abu-Oda and El-Halees [2015],Kabakchieva [2013],Cortez and Silva [2008],Bydžovská [2016],Kulkarni and Ade [2014] and Bydžovská [2016]. The accuracy achieved across these studies range between 70% and 90%. The SVM predictive model was employed by Sangodiah et al. [2015], Ajoodha and Jadhav [2019],Cortez and Silva [2008] and Bydžovská [2016]. The accuracy achieved across these studies ranges between 80% and 99%.

The J48 predictive model was was employed by Semeon [2011], Hambali Moshood, Ajoodha and Jadhav [2019], Kabakchieva [2013] and Dekker et al. [2009]. The accuracy achieved across these studies range between 90% and 99%. The Multilayer perceptron was employed by Ajoodha and Jadhav [2019], Sangodiah et al. [2015] and Awad and Ewais [2018]. The accuracy achieved ranges between 80% and 90%. The K-star was employed by Kulkarni and Ade [2014], Ajoodha and Jadhav [2019]. An accuracy of 99% was achieved The Logistic regression was employed by, Soule [2017], Ajoodha and Jadhav [2019] and Rusli et al. [2008]. Accuracy values range betweem 80% and 82.

3. Methodology

The task at hand as predicting the completion of a learner's Computer Science degree based on their first-year marks. In simpler terms, we are trying to answer the following question, Will a student complete their computer science degree looking at their first year marks? The are three possible outcomes: "Complete", the student is expected to complete their degree in minimum time (3 years); "Fail", the student is expected not to complete their degree; and "Delayed", the student is expected to successfully complete their degree but not in minimum time. Several machine learning predictive classification models to deduce a learner into one of three categories "Completed","Delayed" or "Failed" will be employed.

The best performing algorithm will be used to create an application that will be used to predict the outcome of a student. To gauge the performance of the trained classification models, confusion matrices and AUROC curves will be used.

This section is structured as follows: Firstly, a brief description of the data collection procedure is given. This includes information concerning ethics clearance. Secondly, preprocessing steps taken to prepare the data for this research objective are outlined. Thirdly, a brief descriptions of each machine learning classifier to be used is given . Fourthly, the feature analysis process is presented and finally brief descriptions of the evaluation metrics used to gauge the performance of the machine learning predictive classification models.

A. Ethics clearance and Data collection

Data utilized in this study was obtained from a South African University. The data consists of biographical, student

TABLE I: Systematic Literature Review

Study	Study Purpose	Attributes	Methods/Techniques	Relevant Findings
Ajoodha and Jadhav [2019]	Identify at risk-students	Biographical data	Logistic Regression, MLP, NB, SVM, K*, Random forest	The LG model achieved an accuracy of 83.622%. Grades can be used as predictor variables.
Pojon [2017]	Predict performance of students	Age, Gender, Grades	Linear regression, Decision tress, naive Bayes' classifier	NB classifier achieved the highest accuracy of 95%. Feature engineering increases model accuracy.
Semeon [2011]	Investigate and predict students' dropout	Grades, Sex, Age, Income	OneR,J48, RandomForest, Multilayer Perceptron	J48 algorithm achieved an accuracy of 91%, MLP achieved an accuracy of 89%. Grades can be used as predictor variables.
Sangodiah et al. [2015]	Investigate and minimize students' attrition	1st-yearGrades, age, gender, race, sex, qualification result, current GPA, current CGPA	SVM, Decision Tree	SVM algorithm achieved and accuracy of 86% and DT an accuracy of 81%. Grades can be used as a predictor variable
Abu-Oda and El-Halees [2015]	Predict students' dropout	Grades	Decision Tree, Naive Bayes	Naive Bayes achieved an accuracy of 98% and the Decision tree achieve an accuracy 99%. Grades can be used a predictor variable
Hambali Moshood	predict Students' performance in Computer Programming	Grades	CART, J48, BF tree	Grades can be used as a predictor variable. The J48 achieved an accuracy of 70%. Maths intensive subjects are strong predictors.
Kabakchieva [2013]	Predicting Student Performance	Grades, gender, age,	C4.5 decision tree, J48 ,Naive Bayes, k-NN,OneR	SVM can be used to predict students' performance. It out performed the other algorithms.
Dekker et al. [2009]	Predicting Students Drop Out	Grades, Biographical attributes,	CART, J48, OneR, BayesNet,OneR	J48 achieved an accuracy of 80%. Grades can be used as predictor variables.
Butcher and Muth [1985]	Predict students' performance in computer science course predict final grades	Grades	Statistical system (SAS)	Historic grades can be used to predict students' performance
Cortez and Silva [2008]	predict student performance at a secondary school	Biographical data, Grades	SVM, Naive Bayes	The Naive bayes predictive model outperforms the SVM predictive model
Bydžovská [2016]	A Compare predictive models for student performance prediction	Biographical data, Social Behavior Data, Grades	SVM, Random forests, Naive Bayes, J48, OneR, Baseline	SVM reached the best results.
Kulkarni and Ade [2014]	Incremental learning for students performance	Grades,Biographical, Social data	K*, NNGe, IBK, Naive Bayes	SVM can be used to predict students' performance. It out performed the other algorithms.
Awad and Ewais [2018]	Predict High School Exam Result	Grades	Multilayer Perceptron	The MLP achieved an accuracy of 99%
Soule [2017]	Predict Student performance	Grades	Logistic Regression	The LG model achieved an accuracy of 82%. Grades can be used as predictor variables.
Rusli et al. [2008]	Predict Students' academic performance	Grades	Logistic Regression, Artificial NeuralNetwork ,Neuro-Fuzzy	The LG model achieved an accuracy of 83.622%. Grades can be used as predictor variables.

marks from first year to third , all postgraduate marks and enrollment observations of students from the Faculty of Science doing Mathematical Science Degrees. The study participants are students who studied anytime between the years 2008 to 2017 at a research-concentrated South African university. The Human Research Ethics Committee at the institution approved this study.The committee also addressed ethical issues of protecting the identity of the students participating in the research and ensuring the security of data.

B. Data preparation

First year Computer Science students undertake three major courses and one elective. The data contains marks obtained by each student during their first year. Using this data, the features that will help us predict the completion of a degree were engineered. These features are MajorOne, MajorTwo, MajorThree, Elective, Aggregate and ProgressOutcome. The first four attributes are the predictors and Progress outcome is the target attribute. The target attribute contains three

Variables	Description	Type	Variable encoding values
Major One	Aggregate of final Calculus 1 and Algebra 1 marks	Nominal	4 = 70% - 100% 3 = 60% - 69 % 2 = 50 % - 59% 1 = 0% - 49%
Major Two	Aggregate of all Coms 1 marks	Nominal	4 = 70% -100% 3 = 60% - 69% 2 = 50% - 59% 1 = 0% - 49%
Major Three	Computation and Applied Mathematics orInformation Systems Final Marks	Nominal	4 = 70% - 100% 3 = 60% - 69% 2 = 50% - 59% 1 = 0% - 49%
Elective	Marks of any Level 1 course (i.e. Physics)	Nominal	4 = 70% - 100% 3 = 60% - 69% 2 = 50% - 59% 1 = 0% - 49%
Aggregate	Aggregate of all marks obtained	Nominal	4 = 70% - 100% 3 = 60% - 69% 2 = 50% - 59% 1 = 0% - 49%
Progress Outcome	Outcome of a student in their final year	Categorical	Completed - Yes Failed - No Delayed - NRT

TABLE II: The Students data set description

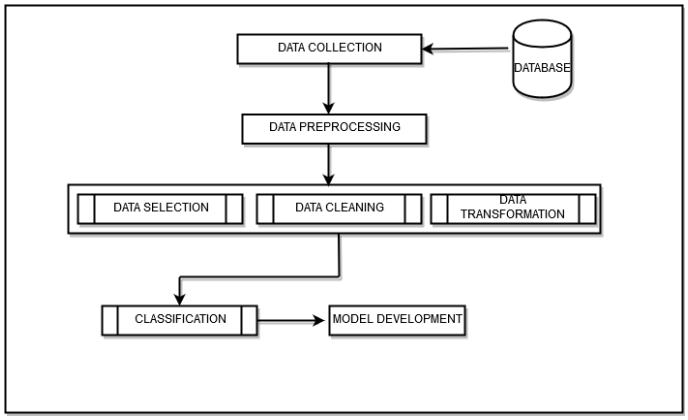


Fig. 3: Proposed Methodology of Classification Model

classes: Completed , Delayed, and Failed. Table III shows the description of these attributes.

C. Data preprocessing

Some entries in the data contained missing values. This is due to a number of reasons. The most common reason being that some students did not enrol for some of the courses. To combat this problem, entries with missing values where removed.The data set contains a total number of 624 instances. According to Fig 4, there is a huge gap between the number of students who are labeled as 'Failed" and those those who are labelled "Completed" or "delayed. The classes are imbalanced , this may lead to reduced accuracy. To combat this, we under sampled the data using the spreadsubsample filter on WEKA. The final data contains 393 instances with 131 instances in each class. The final dataset looks like the Table III. We have five predictors and one target variable.

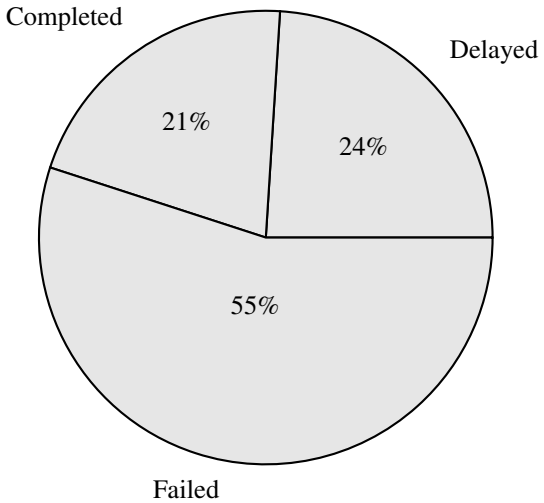


Fig. 4: The percentage of Delayed, Completed and Failed instances contained in the data set

Major One	Major two	Major three	Elective	Aggregate	Outcome
3	3	4	3	3	Completed
2	1	3	2	2	Failed
3	3	4	4	3	Completed
4	2	3	2	3	Delayed
3	4	2	2	4	Delayed
1	1	1	2	1	Failed

TABLE III: The resulting data set after preprocessing

D. Attribute ranking

The goal of attribute ranking is to determine the predictive power of each variable in our feature set. To achieve this, the Information Gain Ranking algorithm will be employed in order to rank our predictor variables. The IGR algorithm calculates the information gain for each feature with respect to target feature [Ajoodha and Jadhav 2019]. Information gain measures how much information a feature gives us about a class. The values of Information gain (entropy) are between 0 and 1, that is, the minimum entropy is 0 and the maximum entropy is 1 [Ajoodha and Jadhav 2019].

E. Classification Algorithms

In this study, the machine learning algorithms employed for the classification process were K-Star, Naive Bayes, SVM, Decision tree, Logistic Regression, and the Multi-layer Perceptron. These algorithms were chosen because they each represent a different class of machine learning algorithms.(i.e) The J48 comes from the "Tree" family and the Logistic regression comes from the Regression family.

K*: The K* instance-based classifier uses an entropy-based distance function to classify test instances using the training instance most similar to them [Ajoodha and Jadhav 2019]. The implementation of the of this algorithm follows the implementation by [Ajoodha and Jadhav 2019].

Naive Bayes: For prediction problems, this is the most used algorithm [Pojon 2017]. It is beloved for its pragmatic approach to machine learning problems. It uses Bayes' theorem to classify instances to one or a number of independent classes using probabilistic approach [Koller et al. 2009]. It is the easiest learning algorithm to implement [Pojon 2017]. It attempts to find the likelihood of features occurring in each class and takes the class with the largest posterior probability as our predicted class. The main assumption is that all features are conditionally independent given the class label of each instance.

SVM: A supervised ML algorithm that finds a multi-dimensional hyperplane that will best divide the dataset into two classes [Pojon 2017]. Test instances are then mapped on the same space and predicted based on which side of the hyperplane they fall on. The use of the Kernel trick enables SVMs to be scaled for nonlinear and classification problems of high dimensions. It was initially developed for binary classification cases but it can be extended to multi-class problems by breaking them to binary class problems[Pojon 2017].

J48: The J48 algorithm , a successor of ID3 was developed bY Ross Quinlan and is implemented in WEKA using Java [Bashir and Chachoo 2017]. The decision tree's greedy top-down approach is adopted by this algorithm. It is used for classification in which new instance is labelled according to the training data.

Logistic Regression: The Linear Logistic Regression model predicts probabilities directly by using the Logit transform. The implementation in this paper follows Sumner et al. [2005].

Multi-layer Perceptron: The MLP used in this paper is a feed-forward neural network that uses the sigmoid functions as activation functions. The sigmoid function is used to represent nodes and back-propagation to classify instances. The implementation in this paper follows Ajoodha and Jadhav [2019].

F. Model Evaluation

To measure the performance of the trained classification models, confusion matrices and Receiver Operating Characteristic curves will be used. From the confusion we will derive performance measure metrics from like Precision, Recall and Accuracy and from the ROC curve we will derive the AUC (area under the ROC curve metric). For all models, a 10-fold cross validation scheme was used [Fushiki 2011]

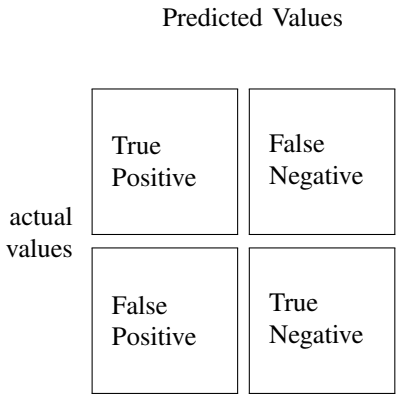


Fig. 5: Confusion Matrix

From the confusion matrix depicted in Fig 5, True Positive gives the proportion of positive entries that were accurately

identified, False Positive is the proportion of positive entries that were classified as negative, False Negative is the fraction of negative entries that were classified as positive and True Negative is the number of negative entries that are classified as negative.

True Positive can be abbreviated as TP, False Positive as FP, True negative as TN and false negative as FN. From Pojon [2017], we define the three metrics as follows Accuracy which is the proportion of correct predictions is

$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN} \quad (1)$$

The higher the accuracy, the better the model's performance. Precision which is the capacity of a model not to classify a positive instance as negative.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

The higher the precision, the better the model's performance. Recall which is ability of a classifier to find all positive instances is

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

The higher the Recall, the better the model's performance. A Receiver Operating Characteristic (ROC) is a curve with that shows the relationship between true positive rates and false positive rates of a model [Lobo et al. 2008]. It shows the relationship between sensitivity and specificity of a model. The area under the ROC curve tells us about the classification model's discriminatory capabilities, that is, the ability of a model to discriminate between classes. The values of the area under the ROC curve (AUC) of a model range between 0 and 1. The higher the AUC of a model, the better the performance of the model.

4. Results and Discussion

In this paper, Six classification models were employed with the purpose of predicting whether a learner will complete their Computer Science degree or not. Learners are categorized into three class, "Completed", "Delayed" and "Failed". Classification models are evaluated using Accuracy, Precision, Recall, AUC and Execution time.

A. Feature Ranking

One important part of the study is feature ranking, whereby the predictor attributes are ranked based on the strength of their relation with the dependant attribute. Information gain can be used to rank the predictor variables according to the strength of their relation with the outcome variable [Ajoodha and Jadhav 2019]. To rank the predictor variables in terms of their predictive power, the IGR algorithm was used. Table IV depicts the ranking of the features using information gain. "Aggregate" has the highest gain ratio of 0.308. "Elective" has the lowest gain ratio of 0.104. Fig 6 depicts a graphical illustration of the information gain for a set of features used to classify a learner as Completed, "Delayed" or "Failed".

Feature	Entropy	Rank
Aggregate	0.308	1
Elective	0.104	5
MajorOne	0.294	2
MajorTwo	0.189	3
MajorThree	0.186	4

TABLE IV: A ranking of the information gain (entropy) for a set of features to predict the class variable (leaner's outcome)

The aggregate feature is ranked as number one because it contains information from all the attributes in the dataset. The attribute ranking also shows that students who do well in MajorOne courses, that is, students who perform well in Calculus and Algebra are more likely to complete their degree. Computer Science is a calculation intensive course and it requires very sharp brains that can think very fast.

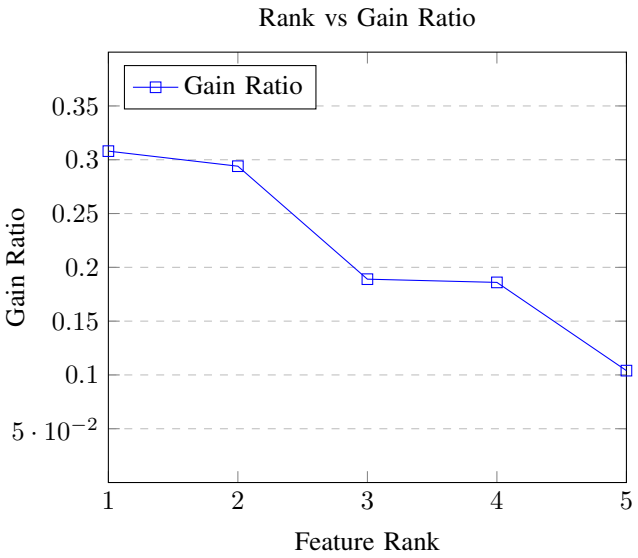


Fig. 6: A graphical illustration of the information gain (entropy) for a set of features to predict the learner's completion outcome (class variable). The x-axis indicates the feature rank and the y-axis indicates the information gain for using that feature.

B. Classification and Model Evaluation

Different classifiers whose results vary depending on efficiency were employed. In this paper the following algorithms were employed: Naive Bayes, SVMs, Decision trees, K* , Multilayer- perceptron and Logistic Regression. To gauge the performance of these classifiers, we used accuracy, precision, recall and AUC Confusion matrices of the predictive classification models are depicted in Figure 7. From the confusion matrices, the accuracy , precision and recall of each model were calculated using 10-fold

cross validation. Figure 7a depicts the confusion matrix of the Naive Bayes predictive model. The model achieves an accuracy of 52%. In terms of precision and recall, the model achieves 0.571 and 0.583 respectively.

Figure 7b depicts the confusion matrix of the Logistic Regression predictive model.The model achieves an accuracy of 60%. In terms of precision and recall, the model achieves 0.605 and 0.606 respectively.

Figure 7c depicts the confusion matrix of the SVM predictive model. The model achieves an accuracy of 58%, In terms of precision and recall, the model achieves 0.581 and 0.580 respectively. Figure 7d depicts the confusion matrix of the K-Star predictive model. The model achieves an accuracy of 57%. In terms of precision and recall, the model achieves 0.580 and 0.571 respectively.

Figure 7e depicts the confusion matrix of the J48 predictive model. The model achieves an accuracy of 70%. In terms of precision and recall, the model achieves 0.69 and 0.69 respectively.

Figure 7f depicts the confusion matrix of the Multilayer Perceptron predictive model. The model achieves an accuracy of 54%. In terms of precision and recall, the model achieves 0.548 and 0.542respectively.
In all the models in Fig 7, the most confusion happens between the failed students and delayed students. The predictive models classifies a number of Failed students as Delayed. This is because there are some students who do badly in their first year of studies but recover and eventually complete their degree. Some students who failed have marks similar to this who recovered and passed.

In addition to these evaluation metrics, execution time (the time the predictive classification model takes to build) was also considered as a measure of the predictive models' efficiency. Table V shows the results obtained. In terms of accuracy, compared to other predictive models on this paper, the J48 predictive classification model outperformed the other predictive classification. It also recorded this highest precision and recall values. The J48 recorded an execution time of 0.09 seconds which is the which is the second lowest execution time recorded on this paper. The K-Star took the least time to build with an execution time of 0.01 seconds. With an exception of the J48 and K-star predictive models, the Naive bayes took the least time to build (0.10 seconds) trailed by the Logistic Regression predictive classification model which recorded an execution time of 0.20 seconds. The Multilayer Perceptron predictive model took the longest time to build (0.56 seconds) whilst the SVM predictive classification model took the second longest time to build (0.40 seconds). Looking at the evaluation metrics derived from the confusion matrix, it is clear that the J48 algorithm performance better as compared to the other 5 predictive classification models in this paper. To verify that this indeed the case, we will use AUC of each predictive model. In Figure 8, the ROC curves of each class along side the micro-average ROC curve of each model are illustrated.

Figure 8a depicts the ROC curves of the K* predictive model. The AUC of "("Completed) class is 0.67. the the Delayed has a value of 0.56 and the "Failed" class and AUC value of 0.84. The micro-average AUC is 0.69. This model predicts well the failed class instances.

Figure 8b depicts the ROC curves of the SVM predictive model. The AUC of "("Completed) class is 0.73. the the Delayed has a value of 0.53 and the "Failed" class and AUC value of 0.86. The micro-average AUC is 0.72. This model predicts well the failed class instances.

Figure 8c depicts the ROC curves of the Multilayer Perceptron predictive model.The AUC of "("Completed) class is 0.64. the the Delayed has a value of 0.56 and the "Failed" class and AUC value of 0.85. The micro-average AUC is 0.68. This model predicts well the failed class instances.

Figure 8d depicts the ROC curves of the Logistic Regression predictive model. The AUC of "("Completed) class is 0.66. the the Delayed has a value of 0.57 and the "Failed" class and AUC value of 0.84. The micro-average AUC is 0.71. This model predicts well the failed class instances.

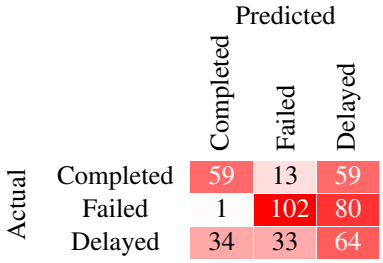
Figure 8e depicts the ROC curves of the Naive Bayes predictive model. The AUC of "("Completed) class is 0.64. the the Delayed has a value of 0.56 and the "Failed" class and AUC value of 0.85. The micro-average AUC is 0.68. This model predicts well the failed class instances.

Figure 8f depicts the ROC curves of the J48 predictive model. The AUC of "("Completed) class is 0.72. the the Delayed has a value of 0.51 and the "Failed" class and AUC value of 0.88. The micro-average AUC is 0.75. This model predicts well the failed class instances.

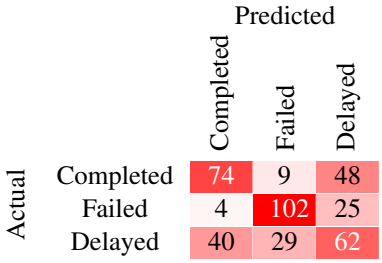
Based on these experiments, the Failed class is most accurately predicted class. The AUC values of the predictive models range between 0.68 and 0.75. The J48 predict attained the highest micro- average AUC of 0.75. It is therefore the best performing predictive model based on the AUC metric. The accuracy level of the predictive models employed in this study range between 52% and 70%. The precision and recall values range between 0.5 and 0.7. For the development of the software application, the J48 Algorithm will be used.

Algorithm	Accuracy (%)	Precision	Recall	Execution time(sec)
Naive Bayes	52.013	0.571	0.583	0.10
SVM	58.015	0.581	0.580	0.40
J48	70.125	0.690	0.692	0.09
Logistic Regresion	60.560	0.605	0.606	0.20
Multi-layer Perceptron	54.199	0.548	0.542	0.56
K-Star	57.506	0.548	0.542	0.01

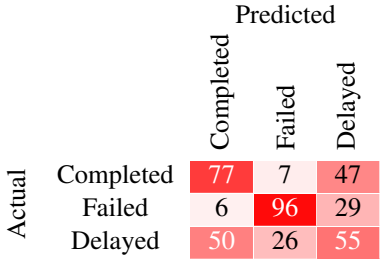
TABLE V: The accuracy , precision and Recall of the 6 predictive classification models used in this paper obtained using using 10-fold cross validation method. The execution time of each model is the time taken by the predictive classification model to build. From The J48 decision tree model outperforms all the models.



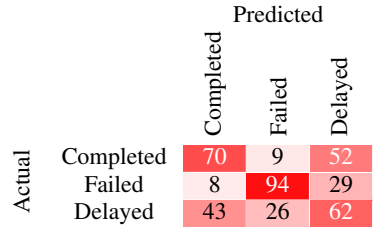
(a) A confusion matrix describing the performance of the Naive Bayes predictive model. The NB model achieves an accuracy of 52%, a precision of 0.58 and a recall of 0.58. Out of the 393 instances, 225 were correctly classified and only 168 were classified incorrectly.



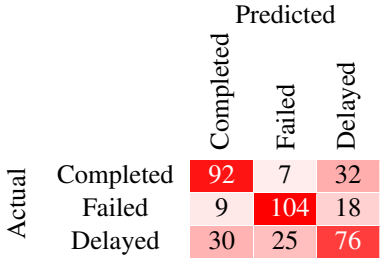
(b) A confusion matrix describing the performance of the Logistic Regression predictive model. The NB model achieves an accuracy of 52%, a precision of 0.58 and a recall of 0.58. Out of the 393 instances, 225 were correctly classified and only 168 were classified incorrectly.



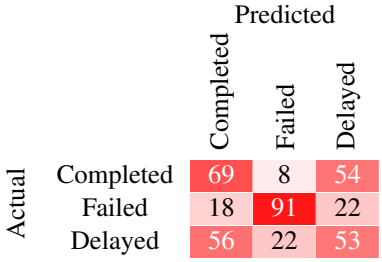
(c) A confusion matrix describing the performance of the SVM predictive model. The SVM model achieves an accuracy of 58%, a precision of 0.60 and a recall of 0.60. Out of the 393 instances, 228 were correctly classified and only 165 were classified incorrectly



(d) A confusion matrix describing the performance of the K-star predictive model. The K-star model achieves an accuracy of 52%, a precision of 0.58 and a recall of 0.58. Out of the 393 instances, 225 were correctly classified and only 168 were classified incorrectly.

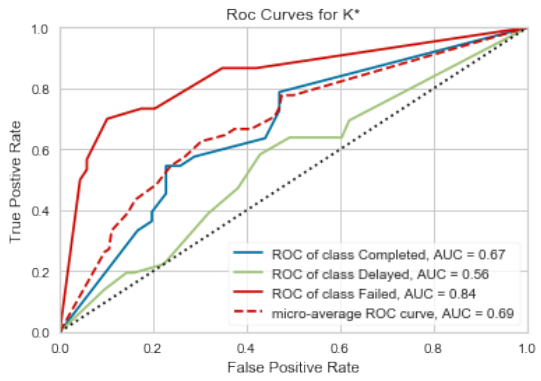


(e) A confusion matrix describing the performance of the J48 predictive model. The J48 model achieves an accuracy of 70%, a precision of 0.69 and a recall of 0.69. Out of the 393 instances, 272 were correctly classified and only 121 were classified incorrectly

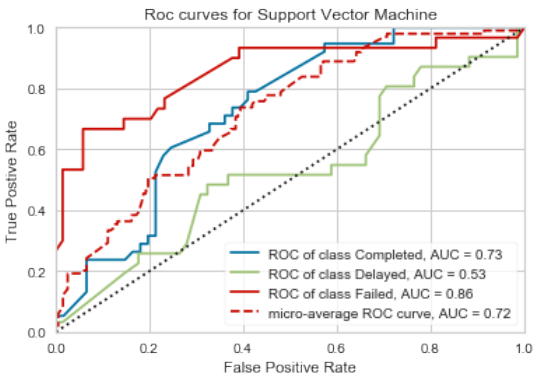


(f) A confusion matrix describing the performance of the Multilayer Perceptron predictive model. The Multilayer perceptron model achieves an accuracy of 52%, a precision of 0.58 and a recall of 0.58. Out of the 393 instances, 225 were correctly classified and only 168 were classified incorrectly.

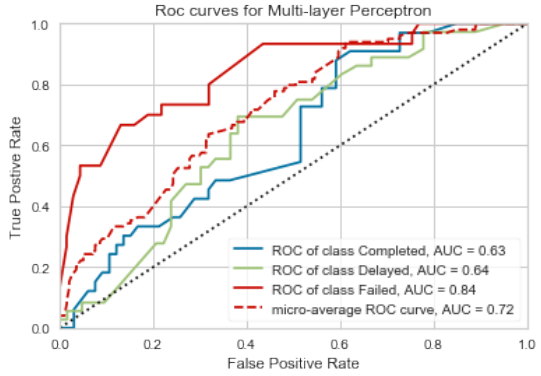
Fig. 7: A set of confusion matrices describing the performance of several predictive models on a set of test data. Each predictive model's accuracy and indicated along with the correctly and incorrectly classified instances



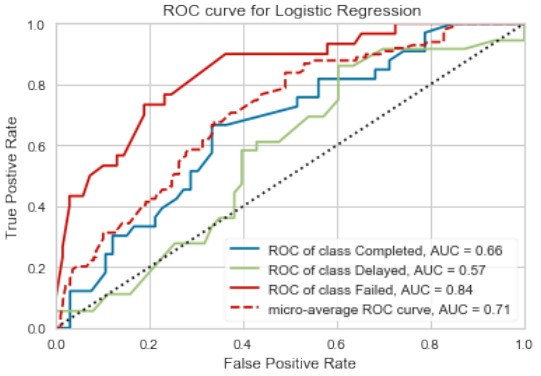
(a) Relationship between sensitivity and the specificity of the **K-star** predictive classification model. The AUC of the **K-star** is 0.69



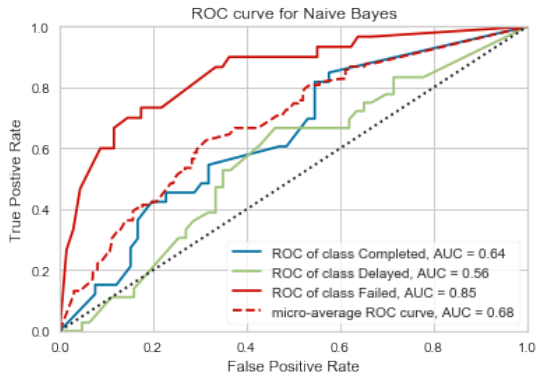
(b) Relationship between sensitivity and the specificity of the **SVM** predictive classification model. The AUC of the **SVM** predictive classification model is 0.72



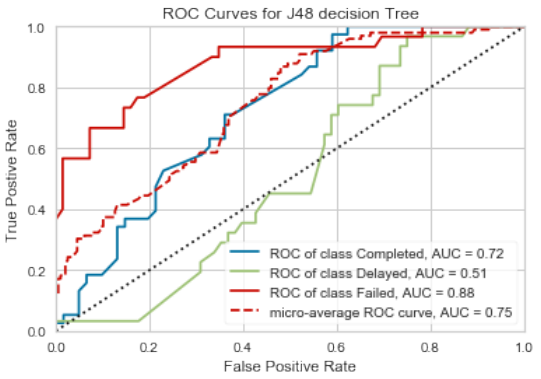
(c) Relationship between sensitivity and the specificity of the **MLP** predictive classification model. The AUC of the **MLP** predictive classification model is 0.72



(d) Relationship between sensitivity and the specificity of the **Logistic Regression** predictive classification model. The AUC of the **Logistic Regression** predictive classification model is 0.71



(e) Relationship between sensitivity and the specificity of the **Naive Bayes** predictive classification model. The AUC of the **Naive Bayes** predictive classification model is 0.68



(f) Relationship between sensitivity and the specificity of the **J48** predictive classification model. The AUC of the **J48** predictive classification model is 0.75

Fig. 8: Figures illustrating the ROC curves of the predictive classification models. AUC values of each class and the average AUC value of the predictive classification models are given.

5. Conclusion And Future Work

The early detection of students in danger of failing to complete their studies is crucial for HEIs. The ability to detect at risk students at an early stage can help HEIs save huge amounts of money and also help to fill the knowledge gap in these institutions. Students at risk of dropping out may be provided with help with the aim of increasing their academic performance. Early discovery of students at risk may also help to reduce academic attrition rates and increase academic retention rates. In this paper, taking a conceptual framework proposed by Spady [1970] as our rationale , we attempted to predict whether a student will complete their computer science degree based on their first-year marks. To achieve this, machine learning algorithms were employed. With a goal of finding the optimal predictive model for this task, we compared the performance of these predictive models using accuracy, precision, recall and AUC. Based on the experimental results, it was found the accuracy level of the predictive models employed in this study range between 52% and 70%. The precision and recall values range between 0.5 and 0.7. The J48 predictive model attained the best accuracy value of 70%, the best precision of 0.69 and best recall of 0.692 . The micro- average AUC values of the predictive models range between 0.68 and 0.75. The J48 predictive model attained the highest micro- average AUC of 0.75. The J48 was therefore chosen as the best performing classifier based on these results. In this study we also focused on feature ranking. To rank features in our dataset, we employed the information gain criteria where it was found that the aggregate variable is the strongest predictor. This makes sense because this variable contains information of all the other variables. With an exception to the aggregate attribute, MajorOne was the highest ranking feature. This implies that students who perform well in Mathematical subjects are more likely to successfully complete their computer science degree.

One of the major contribution is an application software that can predict if a learner will complete their degree, fail to complete or complete after 3 years. The underlying model used to perform the prediction task is the J48 decision tree. Fig 9 serves as an example of how the program works. A student got the following averaged marks: 67% in Maths (MajorOne); 82% in Applied Mathematics (MajorTwo); 75% for COMS (MajorThree); 68% in Physic (Elective); and has a total Aggregate of 73%. The student is classified as Completed, meaning that the student will complete their degree hypothetically.

This study is limited in the following ways:(a) marks are discretized, information on students marks is lost. For example, a learner who got 60% is seen as a student who go 70%;(b) During data preprocessing, null entries were removed, as a result, crucial information might have been lost; the study is limited to computer science students.

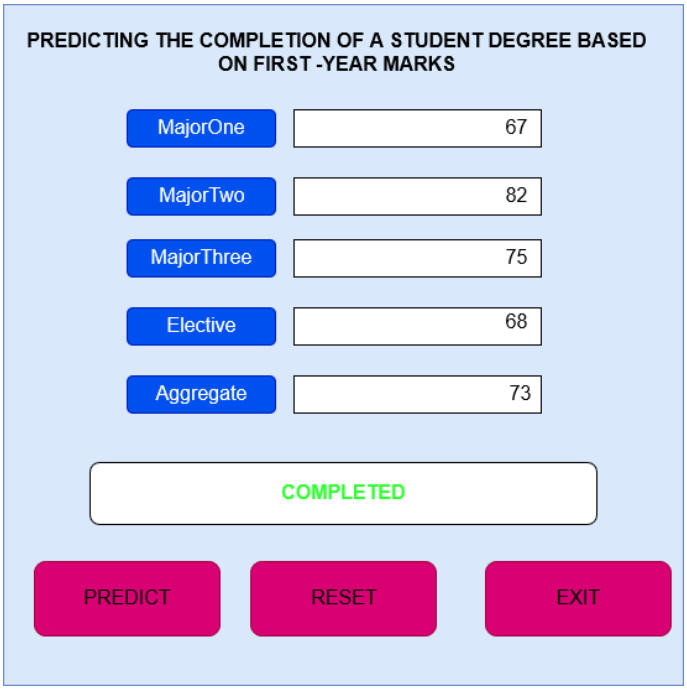


Fig. 9: Application software

For future work, one can: (a) explore the impact of other contributing factors to students performance proposed by Spady [1970],(b) consider including students from other faculties, (c) change the dataset and use numerical values instead of discretized values to avoid information loss.

References

Ghadeer S Abu-Oda and Alaa M El-Halees. Data mining in higher education: university student dropout case study. *Data mining in higher education: university student dropout case study*, 5(1), 2015.

Ritesh Ajoodha and Ashwini Jadhav. Identifying at-risk undergraduate students using biographical and enrollment observations for mathematical science degrees at a south african university. *Arctic Journal*, 72(7):42–71, 2019. ISSN 0004-0843.

Othman Aljohani. A comprehensive review of the major studies and theoretical models of student retention in higher education. *Higher education studies*, 6(2):1–18, 2016.

Mohammed Awad and Ahmed Ewais. Prediction of general high school exam result level using multilayer perceptron neural networks. *Int. J. Appl. Eng. Res*, 13(10):7621–7630, 2018.

Uzair Bashir and Manzoor Chachoo. Performance evaluation of j48 and bayes algorithms for intrusion detection system. *Int. J. Netw. Secur. Its Appl*, 2017.

DF Butcher and WA Muth. Predicting performance in an introductory computer science course. *Communications of the ACM*, 28(3):263–268, 1985.

Hana Bydžovská. A comparative analysis of techniques for

predicting student performance. *International Educational Data Mining Society*, 2016.

Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. 2008.

Gerben W Dekker, Mykola Pechenizkiy, and Jan M Vleeshouwers. Predicting students drop out: A case study. *International Working Group on Educational Data Mining*, 2009.

Tadayoshi Fushiki. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21(2): 137–146, 2011.

A Hambali Moshood. Comparative analysis of decision tree algorithms for predicting undergraduate students' performance in computer programming.

Dorina Kabakchieva. Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, 13(1):61–72, 2013.

Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic graphical models: principles and techniques*. 2009.

Pallavi Kulkarni and Roshani Ade. Prediction of student's performance based on incremental learning. *International Journal of Computer Applications*, 99(14):10–16, 2014.

Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151, 2008.

Edin Osmanbegovic and Mirza Suljic. Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business*, 10(1):3–12, 2012.

Murat Pojon. Using machine learning to predict student performance. Master's thesis, 2017.

Nordaliela Mohd Rusli, Zaidah Ibrahim, and Roziah Mohd Janor. Predicting students' academic achievement: Comparison between logistic regression, artificial neural network, and neuro-fuzzy. In *2008 International Symposium on Information Technology*, volume 1, pages 1–6. IEEE, 2008.

Anbuselvan Sangodiah, PRASHANTH BELEYA, MANORANJITHAM MUNIANDY, LIM EAN HENG, and CHARLES RAMENDRAN SPR. Minimizing student attrition in higher learning institutions in malaysia using support vector machine. *Journal of Theoretical & Applied Information Technology*, 71(3), 2015.

Ian Scott, Nanette Yeld, and Jane Hendry. *Higher education monitor: A case for improving teaching and learning in South African higher education*, volume 6. Council on Higher Education Pretoria, 2007.

Getahun Semeon. Using data mining technique to predict student dropout in st. mary's university college: Its implication to quality of education. 2011.

Patrick Soule. Predicting student success: A logistic regression analysis of data from multiple siu-c courses. 2017.

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

Marc Sumner, Eibe Frank, and Mark Hall. Speeding up logistic model tree induction. In *European conference on principles of data mining and knowledge discovery*, pages 675–683. Springer, 2005.

Surjeet Kumar Yadav, Brijesh Bharadwaj, and Saurabh Pal. Mining education data to predict student's retention: a comparative study. *arXiv preprint arXiv:1203.2987*, 2012.