A Theoretical Model to Predict Undergraduate Learner Attrition using Background, Individual, and Schooling Attributes

Noluthando Mngadi

Supervisor(s):
Dr. Ritesh Ajoodha
Dr. Ashwini Jadhav

A research report submitted in partial fulfillment of the requirements for the degree of Master of Science in the field of e-Science

in the

School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg

7 August 2020

Declaration

I, Noluthando Mngadi, declare that this research report is my own, unaided work. It is being submitted for the degree of Master of Science in the field of e-Science at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.

Noluthando Mngadi
7 August 2020

Abstract

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

Predictive machine learning methods, such as support vector machines, decision trees, and logistic regression, were trained to deduce the students into four risk-profiles. A random forest outperformed other classifiers in predicting at-risk student profiles with an accuracy of 85%, kappa statistic of 0.7, and an AUC of 0.95.

This research argues for a more complex view of predicting vulnerable learners by including the student's background, individual, and schooling attributes.

Keywords: Attrition, At-risk, Machine learning

Acknowledgements

I would first like to thank the one above all, God, if it was not for his love and mercy I would not be where I am today. And I extend my gratitude to my supervisor, Dr Ritesh Ajoodha and my co-supervisor, Dr Ashwini Jadhav, for their dedication and patience, and their time. I would like to thank the financial support from DST-CSIR National e-Science Postgraduate Teaching and Training Platform (NEPTTP). Lastly, the support from my family and friends, their encouragement to arise above everything.

Contents

List of Figures

List of Tables

List of Abbreviations

BN	Bayesian Networks
DT	Decision Trees
ML	Machine Learning
RF	Random Forests
BSc	Bachelor of Science
KNN	K-Nearest Neighbourhood
NB	Naive Bayes
NN	Neural Networks
CPD	Conditional Probability Distribution
NBTQ	National Benchmark Tests
CART	Classification and Regression Trees
ID3	Iterative Dichotomiser 3
XGBoost	Extreme Gradient Boosting Decistion Trees
SVM	Support Vector Machines
IG	Information Gain
AUC	Area Under the Curve
ROC	Receiver Operating Characteristics

To my son Kuhlekonke Shinga (Noncengwa); my late mom Bongisile (Qome) Mngadi; my late uncle Maqhoboza, my day one Ntokozo Masindane; My pillar of strength MaMthalane; my grandmother MaMzwendaba; and the rest of the Ngema family. Ngithi nje, We did it !!!

Chapter 1

Introduction

In this first chapter, we introduce our research. We will look at the background, the problem statement, the motivation of the study, outline the aims and objectives and, finally identify the limitations and assumptions of this research.

1.1 Background

Many students struggle to complete school on time, due to various reasons or factors. For the education institutions to address this issue, they will need to intervene in order to assist students. Intervention can be in the form of support programs that will be applied to assist students in getting back on track or cope at academic institutions. With particular reference to the first years, as they are the most affected Kuzilek et al. [24] and Johnson et al. [21]. To best apply the intervention programs, universities need to identify the contributing factors to risk status, and also identify vulnerable students as early as possible and refer them to appropriate intervention programs.

The report by Bongekile Macupe [27], in Appendix A figure A.1, showed the completion rates of students who enrolled for first year studies in 2013. Rhodes university is leading with completion rate of 73.6%, followed by university of Zululand with 71.6%; and the higher education institution with least completion rate is the university of South Africa with 16.5%. These completion rates are across all university faculties.

Identifying vulnerable (at risk) student profiles is of great benefit to the student, the lecturer, the university, and the government or funders. Studies define at risk concepts differently. Many studies define at risk students as those who drop out or leave before the programs enrolled for complete [8]. This study will adopt the definition of vulnerable learners as those whose interrelations of biographical, individual, and schooling characteristics have a higher probability of failing to meet the minimum requirements to obtain an undergraduate degree in record time (3-years) [1, 5].

A student who enrolls in a higher education program can fall into any of the four risk profiles that are associated with the probability of completing their program. The Risk profiles: 'Lowest risk' - where the student is expected to complete their degree in the minimum time (3 years); 'medium risk' - where the student is expected to complete in more than the minimum time; 'high risk' - where the student fails or drops-out before the minimum time; and 'highest risk' - where the student fails in more than the minimum time.

1.2 Problem Statement (or Research Hypothesis)

In a report by Michael and Susan Dell Foundation in 2017, about 32% of students who are financially supported complete their degree after five years [30]. A spokesperson at the University of Capetown agrees to the fact that some courses experience different rates of dropout in their students. Students in Bachelor of Social Science have the highest dropout occurrences; this is followed by Bachelor of Commerce, while the least occurrence is in Bachelor of Business Science.

Research Questions :

- Is student attrition (risk profile) affected by the background, individual, and schooling attributes ?
- Can we apply machine learning classification algorithms to predict (classify) student risk profiles using background, individual, and schooling characteristics ?
- Which model better or correctly classifies risk profiles ?

1.3 Research Aims and Objectives

This section identifies the aims and objectives of this study. It covers the main goals that we are trying to achieve and how we are going to achieve them.

1.3.1 Research Aims

The main aim of this study is to explore the relationship between background, individual, and schooling characteristics on learner attrition, as per learner attrition model proposed by Tinto [42]. These attributes are then used as input variables to predict student attrition by classifying a learner into four risk profiles: 'Lowest risk', 'Medium risk', 'High risk', and 'Highest risk', as described on Background section.

This study will also compare the predictive accuracy of the Machine Learning algorithms. We will identify which algorithm best deduces the student attrition into the four risk statuses.

Finally, we will use the best model, with high predictive accuracy, to deduce the most significant (important) factors in predicting risk status. The Information Gain (IG) measure will be used to rank the feature contribution, from highest to least important.

1.3.2 Objectives

The research objectives help us achieve our aims stated above :

- Obtain student trajectory data with background, individual, and schooling characteristics, then clean and prune it accordingly.
- Research on the background and current studies from literature.
- Train / build predictive classification (ML) models using the combination of background, individual, and schooling characteristics.
- Use background, individual, and schooling characteristics to predict / classify risk profiles.
- Research previously used models from literature.

- Compare results of our models with the literature results.
- Choose the best model, with high predictive power, from our trained models.

1.4 Limitations

- This study is only limited to students who enrolled at university from the period of 2008 to 2018 and registered for any undergraduate Bachelor of Science degree.
- This is a theoretical model, and its results do not refer to the actual student population.

1.5 Assumptions and Definitions

The assumptions that were made during this research are :

- The synthetic, simulated dataset that was was used to train our models represents the actual population of people who enroll at higher education institutions.
- The duration of a degree is exactly 3 years.
- Enrollment programmes are in the science degrees (BSc) only.

1.6 Overview

In this chapter, we have introduced our topic and highlighted the components that will be covered in our study. The following chapters: **Chapter 2** will look at the background and the related work in studying the effect of biographical characteristics, individual attributes, and schooling factors; on student attrition. Basically we look at what work has been done on this field, and what are the findings of those studies; **Chapter 3** will focus on the methods that will be applied to execute our plan to achieve our aims and objectives, i.e., the data that will be used, the models that will be fitted, and how the whole research was executed; **Chapter 4** we present and discuss the significance of our results and findings; and **Chapter 5** we summarise all the findings, contribution of the paper, and the future work.

Chapter 2

Literature Review

This section expands from the introductory background section that addresses the current literature of the stated problem.

The study of student attrition dates back to the early 1900s by the researchers like Tinto [42] and Mwamwenda [32], till recent studies by Ajoodha and Jadhav [4] and Ajoodha, Jadhav, and Dukhan [5]; where the authors explored the factors affecting student academic performance. However, there is a growing demand for more advanced ways of analyzing educational data and incorporate more information. Student performance is an important metric used to track student and institutional goals, both long term, and short term educational goals.

2.1 Student Attrition

The progress of a student at a tertiary institution is determined by their course final mark or grade, which indicates progress to higher courses Downs et al. [17]. In higher education institutions, there are countless factors within and outside of school that affect the performance of students. The factors that came forward are socio-economic and psychological factors Hijazi and Naqvi [19].

In recent years many studies have focused on distinguishing the critical factors in the student factors that ensure success academically. Characteristics, for example, psychological wellness and social abilities, in particular, self-viability, inspiration, frames of mind and conduct, scholarly competence, communication abilities, team effort, participation, and group capacities, are among the significant highlights for the students to strive or cope at university Hijazi and Naqvi [19]. Students with these aptitudes can work viably with others and deal with their studies productively Hijazi and Naqvi [19] and Lust and Moore [26].

In this research, we adopt the conceptual framework model by Tinto [42], where he relates the background or family, individual attributes, and pre-schooling attributes, to the drop out decisions, Figure 2.1. These features are then used as input to predict student attrition. The combination and relation of these features influence the student's commitment to their goals and school commitment. The input features (i) background or family characteristics, (ii) individual attributes, and (iii) pre-college attribute's impact has been quite explored in previous studies, and provide a right prediction for student performance at higher education institutions.

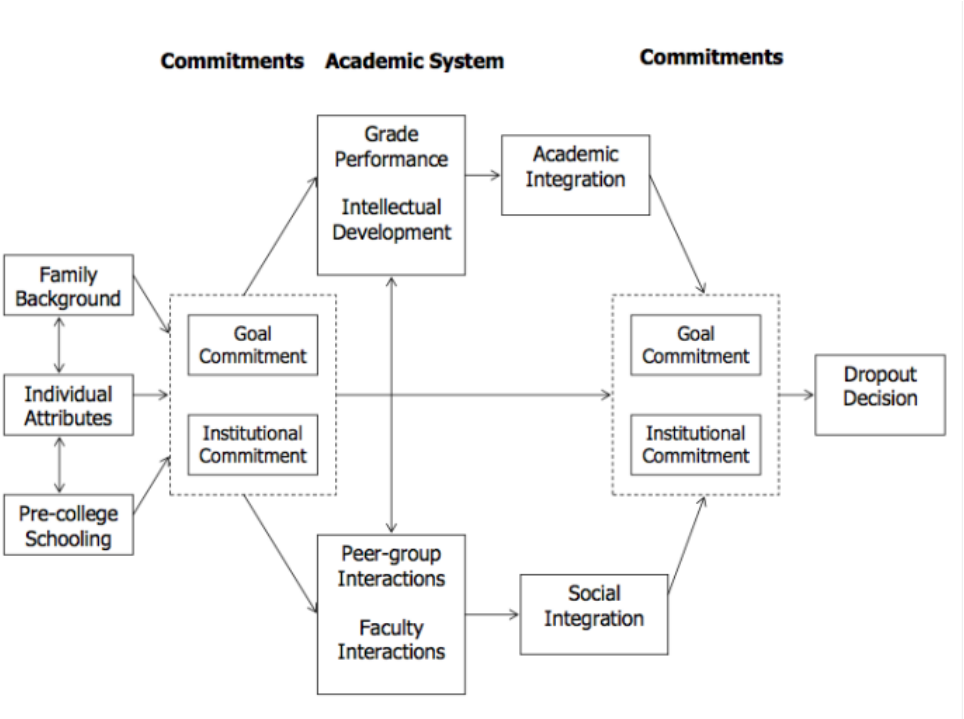


FIGURE 2.1: The Conceptual Framework Model of Tinto [42] that shows the relationship between background or family characteristics, individual attributes, and pre-schooling attributes to the drop out decisions.

Family or Background attributes explored by previous studies include: age, gender, race description, language, family background, living location, parent's occupation and qualifications; to predict the student performance, (Abu Tair and El-Halees [3], Ajoodha and Jadhav [4], Pal [36], Pandey and Pal [37], Downs et al. [17], Mwamwenda [32], Steenkamp, Baard, and Frick [40], Alfan and Othman [6], Yukselturk, Ozekes, and Türel [49], Pandey and Pal [37], Abed, Ajoodha, and Jadhav [1], and Ajoodha, Jadhav, and Dukhan [5]). The findings are quite consistent; gender was found to be a high influencing factor in school drop out, with 68% probability [36]. Similarly, Mwamwenda [32]'s findings showed statistically significant gender anxiety and academic achievement among South African university graduate students. Steenkamp, Baard, and Frick [40] and Downs et al. [17] the English language was found as one of the contributing factors to poor performance, among other factors. More evidence on language issue was discovered by Downs et al. [17], that students from disadvantaged backgrounds (rural areas) perform well on multiple-choice questions, and poorly on essay and short question sections in assessments.

Hard work, self-inspiration, self-viability (effectiveness), student's attitudes and behavior, time the board, and commitment in-class exercises are the variables that fall under individual attributes and contribute significantly to student attrition Stewart [41], Womble [47], and Ajoodha and Jadhav [4]. A large portion of those researches has concentrated on student performance in the U.S. and Europe regions. In any case, since social contrasts may play a role in forming the elements that influence student's performance, it is essential to investigate features according to the region or country Hijazi and Naqvi [19].

In terms of pre-schooling attributes, quite an extensive research has been done on this section by assessing the students' summative assessments (Abu Tair and El-Halees [3], Pal [36], Pandey and Pal [37], Downs et al. [17], Steenkamp, Baard, and Frick [40], Alfan and Othman [6, 7], Kabakchieva [22], and Ajoodha and Jadhav [4]). These factors include entry qualifications and the subjects taken by the student before college. Entry qualifications, and the pre-taken subjects before university show variability in the performance of the students [6]. Pre-taken subjects like economics, mathematics, and accounting subjects are essential in helping the students in choosing the courses in both business and accounting programs [6]. The South African universities solely rely on these attributes (matric results aggregates) as a schooling system measure for acceptance.

2.2 Predictive Modelling in Education

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

In the field of education data mining, many authors have applied a lot of predictive modeling like K-Nearest Neighbourhood (KNN) (Yukselturk, Ozekes, and Türel [49] and Mayilvaganan and Kalpanadevi [29]); Decision Trees (DT) (Yukselturk, Ozekes, and Türel [49], Nghe, Janecek, and Haddawy [33], and Mayilvaganan and Kalpanadevi [29]); Naive Bayes (NB) (Yukselturk, Ozekes, and Türel [49], Pandey and Pal [37], Abu Tair and El-Halees [3], Mayilvaganan and Kalpanadevi [29], and Ajoodha and Jadhav [4]); Neural Networks (NN) (Yukselturk, Ozekes, and Türel [49] and Ajoodha and Jadhav [4]. These models have been widely used for predicting drop out (Yukselturk, Ozekes, and Türel [49] and Bhardwaj and Pal [10]); prediction of student grade Abu Tair and El-Halees [3]; and predicting performer or under performer Pandey and Pal [37].

In terms of relating the features used by previous authors as per the Tinto [42] conceptual framework, i.e. (i) biographical or background, (ii) individual, and (iii) pre-college or schooling attributes; to the machine learning models used to predict learner attrition, by previous studies, and also the model performance.

Nghe, Janecek, and Haddawy [33] extensively researched applying the Bayesian network and decision tree in the forecast of learner's academic behavior. The research reveals that the decision tree performed better than the Bayesian network. However, Kabakchieva [22] discovers that when considering the rates of prediction, the applied data mining algorithm's performances are quite similar. The outcome of the research by Yukselturk, Ozekes, and Türel [49] reveals that k-NN performed more than the other algorithms having a sensitivity of about 87%, decision tree also performed excellently with 79.7%, followed by Neural Network (NN) with 76.8%, while Naive Bayes got 73.9%; on the investigation of the dropout scenario of an online study platform. The Naive Bayes applied by Bhardwaj and Pal [10] achieved excellent outcomes with a dropout precision of 0.917 and a recall of 0.924.

Chapter 3

Research Methodology

In this chapter we will briefly look into the methods that were used to analyse and model our data.

3.1 Research Hypothesis

Background, individual, and schooling attributes can together be used to predict undergraduate learner attrition. This will contribute (produce) to a more complex view of predicting student attrition, rather than only relying on pre-schooling characteristics as a primary schooling system measure for student placement at university or higher education institutions.

3.2 Research Design

In the figure 3.1, we define the pipeline for classifying student risk profiles (attrition). The figure shows the involved stages (phases): data preparation and pruning, feature selection, model training and validation using k-fold cross validation with 10 folds, model testing using test dataset, and model predictions.

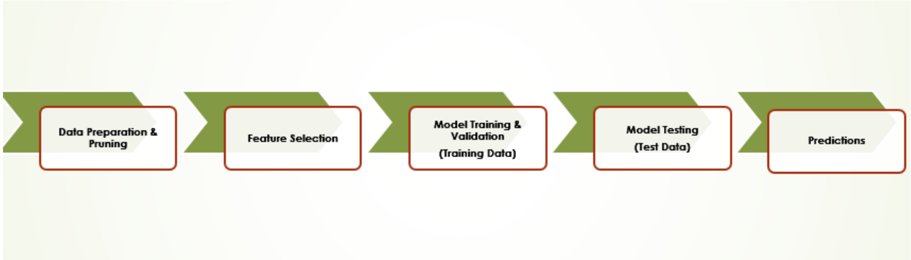

Data Preparation & Pruning
Feature Selection
Model Training & Validation (Training Data)
Model Testing (Test Data)
Predictions

FIGURE 3.1: The research methodology pipeline.

3.3 Data

The data that was used in this study is synthetically simulated from a learned Bayesian Network structure, on the study by Ajoodha and Jadhav [4]. The forward sampling algorithm was applied when generating the data. The values of the parent nodes are sampled from their unconditional distribution, and then the children nodes values are sampled from the parent's sets. The sampling process is iterative until all the nodes values are generated.

The continuous variables were simulated using Gaussian distribution. The Gaussian distribution, well known as Normal distribution, is a continuous function with mean (μ) and standard deviation (σ) and assumes normality in the data. Hence, that is why continuous variables values contained negative values, for example, aggregate marks and probabilities. The negative values are removed from the dataset; we could not modify them as that would temper the distribution of the network. The discrete variables were simulated using tabular conditional probability density (CPD), where you specify the factor levels.

3.3.1 Data Preparation and Pruning

The missing values accounted for about 20% for urban or rural variable and 10% for home province variable. All the observations with missing values were removed entirely on the dataset.

The complete dataset came to 24 variables and 2000 observations, after pruning. We then performed analysis and modeling using the complete dataset, with no missing values.

3.3.2 Theoretical Framework and Features Used

The study adopts the conceptual framework 3.2 to explain the assumption of causal relationships of the variables and the target variable, risk status. The framework hypothetically assumes that the contributing factors to student performance are (i) Biographical characteristics, (ii) Pre-College Observations, and (iii) University Enrolment Observations.

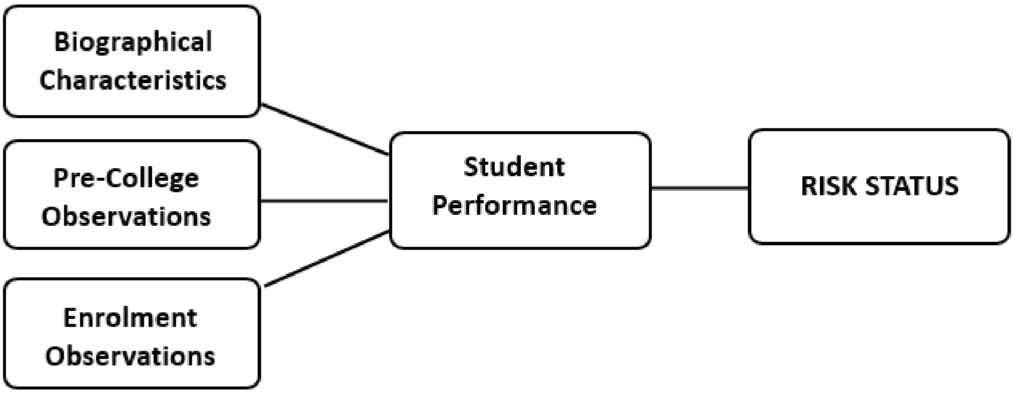


FIGURE 3.2: The theoretical framework.

The table 3.1 summarises the features used under each category. For (i) Biographical characteristics, we have the gender, race of the person, the age at first year, home language, home province, home country - where the person originates from, and whether the person is from rural or urban areas. (ii) Pre-College Observations include school quintile - which indicates school poverty with quintile one is the poorest and quintile five as the least poor school, core mathematics, English first additional language, computer studies, technical mathematics and national benchmark tests (NBTAL, NBTMA, NBTQC) - which measure the student's academic readiness for university. Finally, for (iii) University Enrolment Observations; the year the degree was started, plan description - the professional career, probability of being successful in different science streams (mathematics, physical science, earth science, and biological science), aggregate for course marks, and the number of years spent to complete the degree.

The description of the features and their possible values are attached in Appendix A, table A.1.

TABLE 3.1: List of features used in the study.

Biographical Characteristics	**Pre-College**	**Enrollment Observations**
Gender	School Quintile	Year Started
Race	Mathematics Major	Plan Description
Age at first year	English FAL	Prob of (Math, Physics, Earth, Biological)
Home Language	Computers	Aggregate
Home Province	Additional Maths	Number of years in degree
Home Country	NBTAL, NBTMA, NBTQC	
From Rural / Urban		

3.4 Methods

In this section, we describe the procedure that was applied during this research. We will look into detail the instruments and or software used to carry out modeling, also the methods that were used to model or analyze the data, and finally discuss the metrics that were used to evaluate our trained models.

3.4.1 The Models

In this section, we discuss the predictive models that were used in training our data.

A. Random Forests

Random Forests (RF), also referred to as random decision forests, are an ensemble learning method for classification and regression. They fit several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting, which was the weakness of decision trees [35]. Breiman [11] introduced the bagging method and the randomness of feature selection. The bagging method improved decision trees by lowering the variance and thereby controlling over-fitting. Assuming that we have a training set $X = x_{1,2,\ldots k}$, response variable θ_i , and training tree h_k .

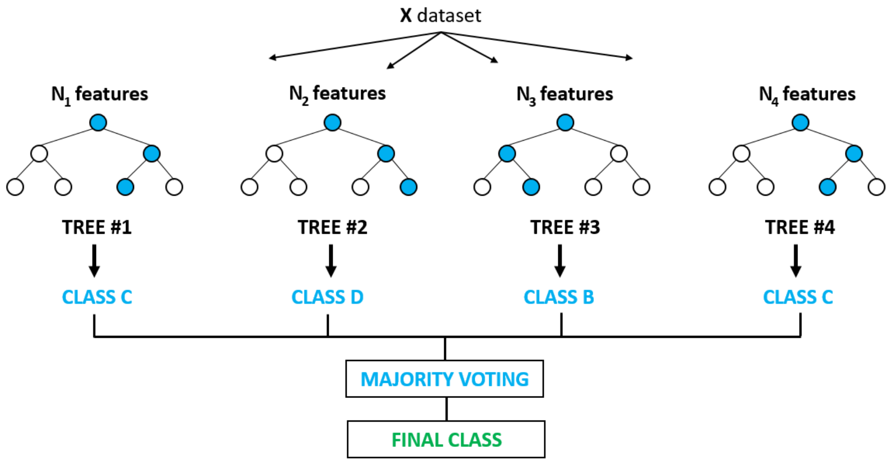


FIGURE 3.3: The Random Forest Example [2].

Figure 3.3 shows an example of a random forest model with X dataset, N features, and the number of trees is 4. The random forest uses the majority class voted by the trees, as the predicted label.

The random forest parameters (in Caret and ranger R packages [23, 48]) are : mtry - "is the number of variables randomly sampled as candidates at each split", split - rule applied at each node splitting, and min.node.size - minimal node size.

B. Coarse Decision Trees

Coarse Decision Trees are the simplest of decision trees. They provide very low model flexibility, only have a few leaves to make a coarse distinction between classes, the maximum number of splits is 4. Decision Trees (DT) are mostly well known as Classification and Regression Trees (CART). DT are tree-like graph models based on possible conditional outcomes. Building a tree requires (involves) decisions on selecting features and deciding the conditions to use for dividing nodes or splitting.

The coarse decision tree parameters (in Caret and rpartScore R packages [23, 18]) are : complexity parameter (cp) - controls over fitting, number of splits, split - split function, and prune - pruning measure.

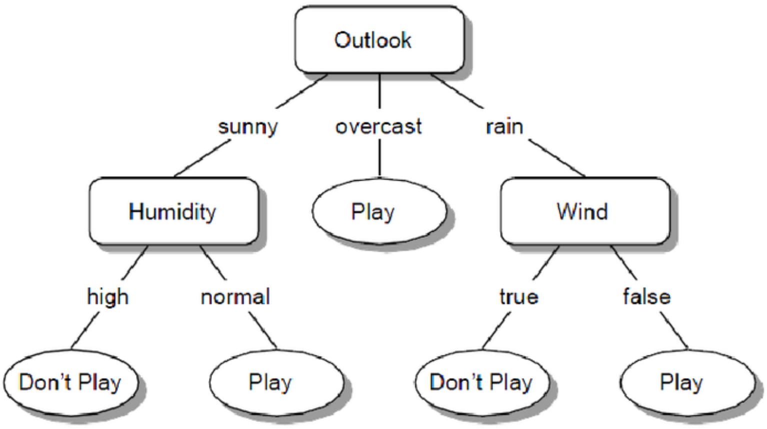


FIGURE 3.4: The Decision Tree Example [15].

C. Linear Logistic Regression

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

The dependent variable Y_i is our risk status with categories: lowest risk, medium risk, high risk, and highest risk; and independent variables X_i's are biographical characteristics, pre-college, and enrollment observations as per table 3.1.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

$$Y_i = \log \frac{Pr(lowest)}{Pr(highest)} \qquad \text{for lowest risk category vs highest risk}$$

$$Y_i = \log \frac{Pr(medium)}{Pr(highest)} \qquad \text{for medium risk category vs highest risk}$$

$$Y_i = \log \frac{Pr(high)}{Pr(highest)} \qquad \text{for high risk category vs highest risk}$$

Where Highest risk category is the reference category

The parameters of the linear logistic classifier (in caret and CaTools R packages [23, 46]) is the number of iterations.

D. J48 (C4.5) Decision Trees

The J48 (C4.5) decision trees is a supervised classification algorithm that is an extension of Quinlan's earlier Iterative Dichotomiser 3 (ID3) algorithm used to generate a decision tree developed by Quinlan [39]. This method is originally applicable in Weka software, but R studio software also provides an RWeka package that allows the implementation of Weka machine learning algorithms. The C4.5 algorithm works well with all data types, continuous, discrete, and text (factor), and it also handles missing observations, meaning it can train with a dataset that has missing values.

Like a typical decision tree algorithm, the C4.5 uses the information gain to select the splitting criteria (attribute), during the tree growth, which is the first step of building (constructing) a tree — followed by the tree pruning, which is the final step of constructing a tree, where the algorithm reduces the error rate by substituting the internal node by a leaf node Podgorelec et al. [38].

The J48 (C4.5) decision trees parameters (in Caret and RWeka R packages [23, 20] are : C - confidence threshold, and M - minimum instance per leaf.

E. Extreme Gradient Boosting Tree

Extreme Gradient Boosting, well known as (XGBoost) is an algorithm that makes use of gradient boosting decision tree algorithms. It computes residuals of prior fitted models, then use this to create new models, that will correct these errors, and thereby improving each new model, until they can no longer be improved. XGBoost applies the gradient descent algorithm to reduce the training error on new models. Therefore this is called gradient boosting [13, 14, 44].

XGBoost is designed for model training efficiency, i.e., train faster and performs better than other tree classifiers, because of gradient boosting. Application and result analysis of XGBoost is elementary and straightforward. They have an advanced functionality of finding the important variables in the model and rank them from the highest to the least important variable and subset the variable list [13, 14, 44].

The Extreme Gradient Boosting decision trees parameters (in Caret and xgboost R packages [23, 14] are : nrounds - boosting iterations, lambda - L2 regularization, alpha - L1 regularization, and eta - learning rate.

F. Support Vector Machines

Support Vector Machines (SVM), are a type of supervised learning algorithm that is applied to both regression and classification problems. They are usually applied to classification problems.

SVM's create a linear line (plane) that separates different (distinct) classes, and that line is called hyperplane. They can be more than one hyperplanes, depending on the number of features used in training the model. The algorithm then finds points that lie closest to the classes, and closest to the hyperplane, the points are called support vectors. The main objective is to find the optimal hyperplane and thereby maximizing the margin, which is the gap between support vectors and hyperplanes.

The Support Vector Machines (SVM) parameters in Caret and e1071 R packages [23, 16] are : kernel, and cost.

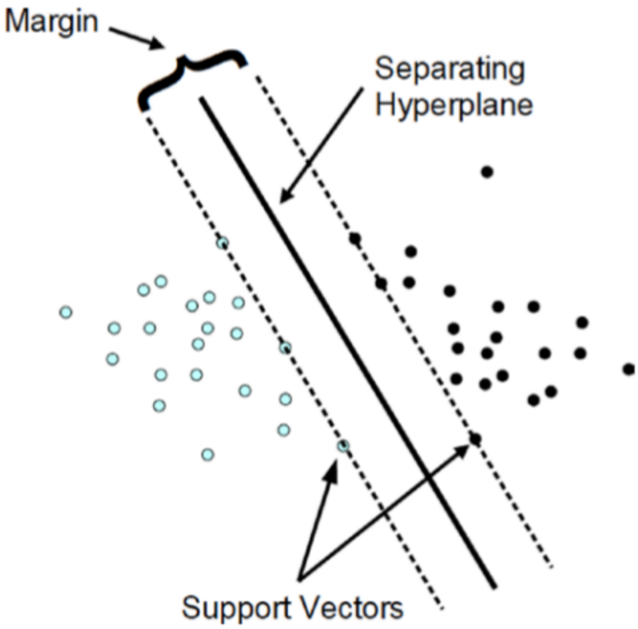


FIGURE 3.5: The example of a Support Vector Machine (SVM) [31].

Figure 3.5 illustrates an example of the application of SVM in classifying two classes: blue circles and black circles. The thick black line in the figure, is the optimal hyperplane, and the support vector points are located (lie) on the lines, and also the margin indicated by the distance between two support vector lines [31]. Support Vector machines are well known for their effectiveness when solving high dimensional problems, and also their memory efficiency.

3.4.2 Model Training

In this section, we look at the smaller components that play a part in model training. We will look at the Instruments and software we used for training our models, and also the decisions and procedures followed.

- **Instruments :** The softwares that we used for modelling our data are: R / R-studio, Weka and Matlab.

  1. R / R-studio is an open source software for statistical computing or programming.
  2. R-studio packages: caret and RWeka for model training; and ggplot2 for graphics.
  3. Waikato Environment for Knowledge Analysis(Weka) is a free open source machine learning software, originally by the University of Waikato, New Zealand.
  4. Matrix laboratory (Matlab) is originally a numeric computing software, but provides toolboxes for machine learning and graphic interfaces.

- **Training/testing split :** Our data was split into two parts, 75% for training our model, and 25% for testing our models. We used splitting by target variable (risk status) into equal class proportions. For example, if class A is 10% in the original dataset, then this split will result in class A proportion of 10% in both training and testing datasets.

- **Model Validation :** For evaluating our model performance, we used k-fold cross-validation, with $k = 10$ folds, where k is the number of groups to split the data. The data is then randomly partitioned into k subsets of equal size.

Each subset is used in turn to validate the model fitted on the remaining k - 1 subset. **A lower expected loss value is better** Tsamardinos, Brown, and Aliferis [45] and Berger [9].

- **Class Imbalance :** The distribution of the target variable, Risk status shows class imbalance, and the High Risk class is most dominant with 67%, followed by Medium Risk with 16%, then Lowest Risk is 9%, and the lowest is Highest Risk with 8%.
  The models were first trained with an imbalanced dataset, then followed by applying the sub-sampling technique, the synthetic minority over-sampling (SMOTE). This method basically over samples the minority class synthetically, and then also under-sample the majority class randomly Chawla et al. [12].
- **Feature Selection :** We first trained our models, then select the best model with the highest predictive accuracy. Then used the best model to deduce the most contributing factors (features) with high Information Gain (IG). A higher IG, when compared to other features, indicates higher importance in prediction. IG scale ranges from zero to one, with zero least contributing and one most contributing (highest IG).

3.5 Analysis

This section describes the tools and metrics used to evaluate the results of the trained models. It provides the baselines that were used to determine the significance of the results.

Metrics

In this section, we describe in detail the metrics that we used to evaluate how well the models fit the data.

A. Confusion Matrix

A confusion matrix is a table that is used to describe the performance of a classification model (or “classifier") on a set of test data for which the true values are known. By known we mean the ‘actual' risk profile is pre-defined, i.e we knew when the student graduated, how long it took to complete the qualification (graduate), and also it was known whether they were successful (qualified) or not. The following algorithm shows how actual risk was defined:

Algorithm 1 : Risk = f (isQualified, numberOfYears)

```
Ensure: $Risk = f(isQualified, numberOfYears)$
  if (isQualified = yes ) then
    if (NumberOfYears = 3 ) then
      Risk = “lowest risk"
    else
      Risk = “medium risk"
    end if
  else
    if (isQualified = no ) then
      if (NumberOfYears < 3 ) then
        Risk = “high risk"
      else
        Risk = “highest risk"
      end if
    end if
  end if
```

TABLE 3.2: An example of a confusion matrix with n classes [28].

Actual Class	Predicted Class			
	Class 1	Class 2	$\cdots$	Class n
Class 1	X_{11}	X_{12}	$\cdots$	X_{1n}
Class 2	X_{21}	X_{22}	$\cdots$	X_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Class n	X_{n1}	X_{n2}	$\cdots$	X_{nn}

Where values for each class i:

$$\text{The True Positive Rate (TP)} = \sum_{j=1}^{n} X_{jj} \tag{3.5.0.1}$$

$$\text{The True Negative Rate (TN)}i = \sum_{\substack{j=1 \\ j\neq i}}^{n} \sum_{\substack{k=1 \\ k\neq i}}^{n} X_{jk} \tag{3.5.0.2}$$

$$\text{The False Negative Rate (FN)}i = \sum_{\substack{j=1 \\ j\neq i}}^{n} X_{ji} \tag{3.5.0.3}$$

$$\text{The False Positive Rate (FP)}i = \sum_{\substack{j=1 \\ j\neq i}}^{n} X_{ij} \tag{3.5.0.4}$$

The precision metric, equation 3.5.0.5 helps us answer the question of when the model predicts the class positively (correct), how often is it correct? In simple terms, this metric gives us the rate at which positive predictions made are correct.

$$\text{The Precision}(i) = \frac{TP_{all}}{TP_{all} + FP_i} \tag{3.5.0.5}$$

The Recall metric equation 3.5.0.6, also known as sensitivity, tells us about the stability of the model in distinguishing classes, which helps us answer the question of how good is our model at detecting the positives. In simple terms, a change in the dataset, how much does it affect the predictions made.

$$\text{The Recall}(i) \ = \frac{TP_{all}}{TP_{all} \ + \ FN_i} \tag{3.5.0.6}$$

The precision and recall are significant metric scores, but it is difficult to maximize both of them, so one has to trade off one metric. However, there is an F - score which is a harmonic mean of precision and recall, one can be able to find the right model, that maximizes F1 score, equation 3.5.0.7.

$$\text{The F Score} \ = 2 \ * \ \frac{\text{Precision} \ * \ \text{Recall}}{\text{Precision} \ + \ \text{Recall}} \tag{3.5.0.7}$$

The accuracy metric is one of most commonly used and important metrics used to measure the performance (correctness) of the model.

$$\text{The Overall Accuracy} \ = \ \frac{TP_{all}}{\text{Total number of test sets (labels)}} \tag{3.5.0.8}$$

C. Kappa Statistic

The kappa statistic is a metric used to asses the classifier performance, especially imbalanced class datasets. This metric compares the agreement between predicted classifications and actual true class labels.

For example, suppose a dataset has two classes, 95% class 1, and 5% class 2 labels. The predictive classifier model achieves 95% accuracy, which in theory, that is a perfect model. However, taking a closer look, the model is only able to detect all class 1 instance but fails to detect (capture) any class 2 observations, that is not a good model, even though the accuracy rate is the best. Hence, that is when the kappa statistic is able to judge the performance of this classifier as not good at all.

The kappa value close to zero means no agreement; the classes are randomly assigned; kappa value close to 1 means close total agreement, and the assignment of classes to of the labels is not random.

The formula :

$$K = \frac{N \sum_{i=1}^{n} m_{i,i} - \sum_{i=1}^{n}(G_i C_i)}{N^2 - \sum_{i=1}^{n}(G_i C_i)} \qquad (3.5.0.9)$$

Where :

i - is the class number.
n - total number of classes. N - is the total number of classified values compared to the true class labels.
m$_{i,i}$ - is the number of values belonging to the actual class i that have also been classified as i (True Positives).
C_i - is the total number of predicted values belonging to class i
G_i - is the total number actual class labels belonging to class i.

Interpretation :
For kappa values interpretation, we will adopt the interpretation table by Landis and Koch [25], for (k) values: $-1 \leq 0 \leq 1$

Kappa (k)	**Interpretation**
< 0	poor or no agreement, random
$0.01 - 0.20$	slight agreement
$0.21 - 0.40$	fair agreement
$0.41 - 0.60$	moderate agreement
$0.61 - 0.80$	substantial agreement
$0.81 - 1.00$	almost prefect agreement

TABLE 3.3: Kappa values Interpretation, according to Landis and Koch [25] recommendations.

D. AUC / ROC Curves

The Area Under the Curve (AUC) - Receiver Operating Characteristics (ROC) curve is a function of sensitivity plotted against 1- specificity at a specific decision threshold. A good model (test), is the one with a ROC curve that passes through the top left corner, where sensitivity and specificity are equal to one. This implies that the ROC curves closer to the top left corner, have a higher overall accuracy of the model [50].

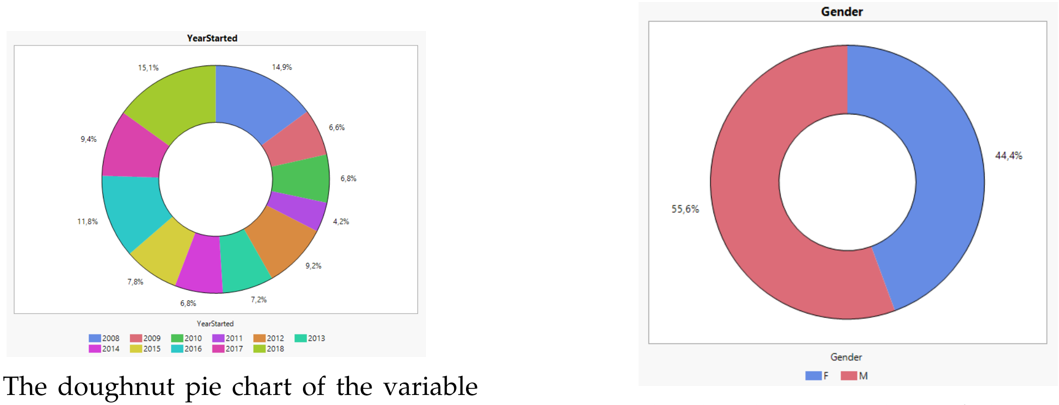
Chapter 4

Results and Discussion

In this chapter, we present and discuss the significance of our results and findings.

4.1 Descriptive Analysis

In this section, we will look at the descriptive statistics, or rather an analysis of the relationship between variables. We also explore how the independent variables relate to each other and the target variable (risk status).

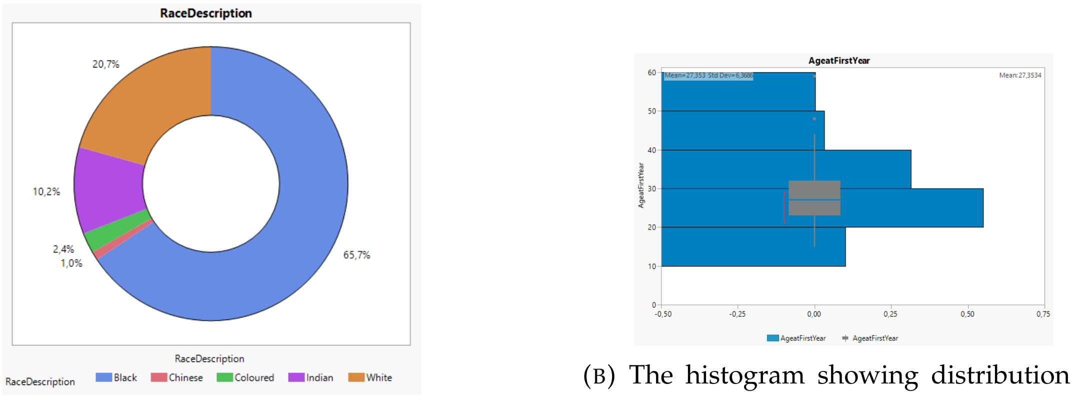


(A) The doughnut pie chart of the variable Year started.

(B) The doughnut pie chart of the gender variable.

FIGURE 4.1: The distribution of the variable year started and that of gender variable.

The figure 4.1 shows the distribution of the variables year started and the gender of the learner. The year started feature ranges from 2008 to 2018, with fairly distributed data over the years, there is no class imbalance. The distribution of gender shows that the proportion of male gender is 55%, and that of females if 44%. The gender variable is evenly distributed.

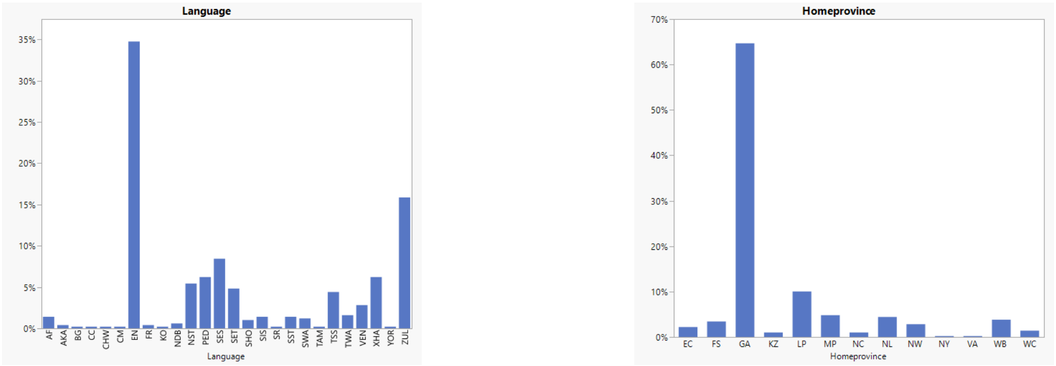


(A) The doughnut pie chart of the race description of the learner.

(B) The histogram showing distribution of the age at first year of the learner.

FIGURE 4.2: The distribution of the race description and the age at first year enrollment.

In figure 4.2, we describe the distribution of the race description and the age of the students at first year enrollment. The proportions of the race descriptions show that the highest proportion is black or African (67%), followed by White (20%), Indian (10%), Coloured (2%), and the least is Chinese (1%). The distribution of age at first year enrollment shows that the age between 20 to 30, is the most common age group for first time enrollment.



(A) The bar graph of the frequency distribution of home language.

(B) The bar graph of the frequency distribution of home province.

FIGURE 4.3: The frequency distributions of the home (native) language and the home province (state) the student originates.

In figure 4.3, we describe the distribution of the home language and home province variables. The majority of the students are English home language (35%), followed by the isiZulu home language speaking (15%), and the rest are around 8% and lower. The majority of the enrolled students originate from the Gauteng province (state) more than 60%, followed by Limpopo province (10%), and the rest are much below 10%.

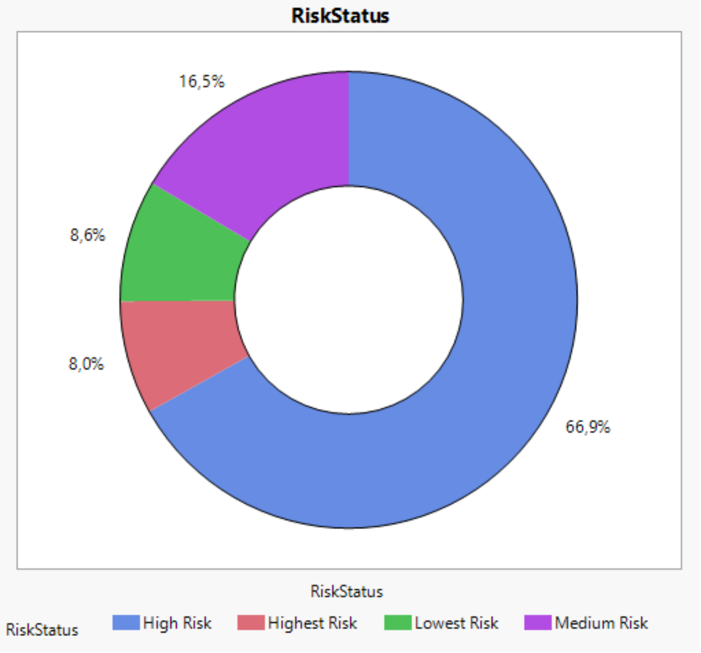


FIGURE 4.4: The distribution of the target variable risk profile (status).

The target variable risk profile has four risk profiles: lowest risk profile, where the student is expected to complete the degree in minimum time (three years); medium risk profile, the student is expected to complete the degree in more than minimum time; the high risk profile, the student is expected to complete the degree after failing and after a long time (maybe 6 years); and the highest risk profile, the student is expected not to complete the degree, i.e., will fail to meet the minimum requirements of the degree and will drop out. Figure 4.4, describes the distribution of our target feature, risk status. It is what our machine learning problem is about. The study aims to classify or deduce the students into the four risk profiles. The classification algorithms are applied to predict this feature or the risk status of the student.

In our dataset, the distribution of the risk profile classes is: the most substantial proportion is high risk profile (67%); followed by medium risk profile (16%); then the lowest risk profile (9%); and the least proportion is the highest risk profile (8%). Hence, this shows that this problem has an imbalanced class dataset. Machine learning algorithms tend to be biassed and favor the high proportion class, which is why there will be an application of class balancing techniques before training our machine learning models.

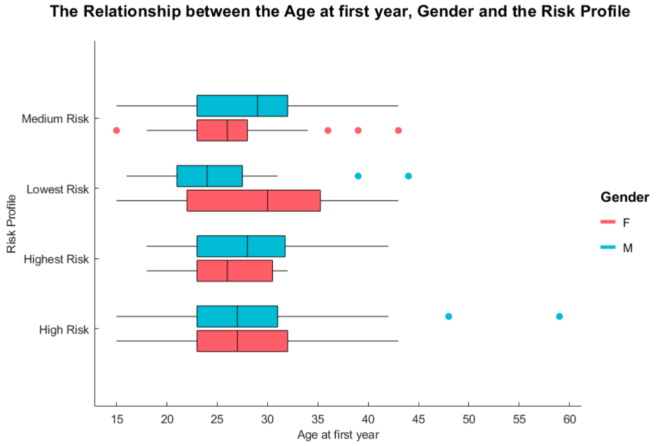


FIGURE 4.5: The relationship between the age at first year, gender and the risk profile.

The average age at first year enrollment at higher education institutions is 27 years old, with a minimum of 15 years and maximum is 45 years, with extreme age of over 50 years. Figure 4.5 describes the relationship between the age at first year, gender, and the risk profile. The distribution of gender shows no much variation with respect to the age of first year enrollment students. However, when looking at the risk profile, there is quite some variation, especially for the lowest risk profile and medium risk profile. For the lowest risk profile, the average age in the first year is 30 for females, and it is less than 25 for males. For the medium risk profile, the average age for females is around 26 years, and for males, its higher than 30 years.

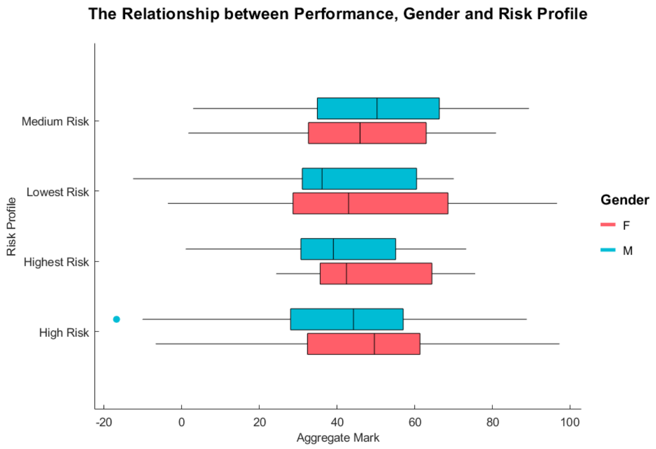


FIGURE 4.6: The relationship between the performance, gender and the risk profile.

The average aggregate mark is 50% across all genders and risk profiles. The variance of performance is very low for both males and females, but females seem to be getting higher aggregates than males for most of the risk profiles.

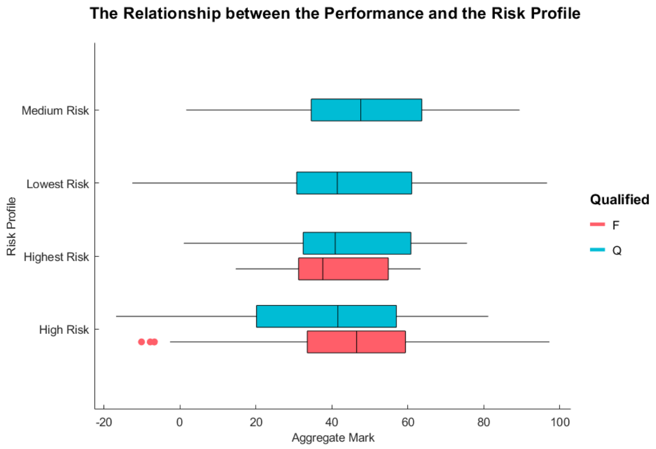


FIGURE 4.7: The relationship between the performance and the risk profile.

The variable qualified has two statuses: qualified and failed, where they refer to whether the student has qualified for the degree or not (which means has failed). Figure 4.7 describes the relationship between the performance and the risk profile. The distribution of qualified significantly differs for different levels of risk profiles.

For high risk profile, there is a higher proportion of failed than qualified; then, for the highest risk profile, the range of aggregate mark is higher for qualified when compared to failed for the same risk profile. The lowest risk profile and the medium risk have no failed learners under their class.

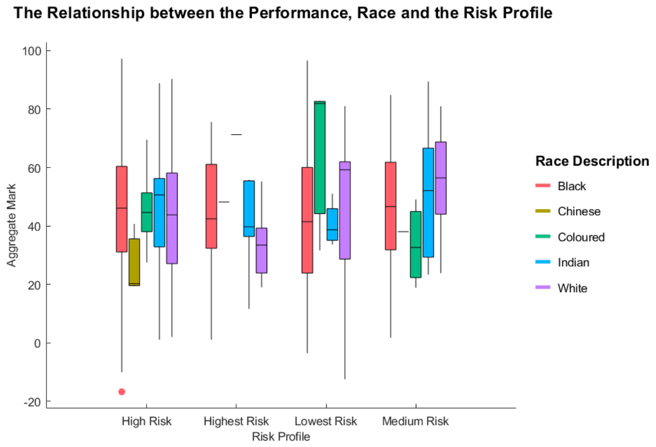


FIGURE 4.8: The relationship between the performance, race and the risk profile.

The study dataset has five race descriptions with their respective proportions: Black (65.7%), Chinese (1%), Coloured (2.4%), Indian (10.2%), and White (20.7%). For the high risk profile and highest risk profile, Blacks perform better than the other races (have higher aggregates); for the lowest risk profile, Coloureds perform way better than other racial groups. The lowest risk profile has an overall higher aggregate for all race descriptions across all risk profiles.

4.2 Feature Importance

This section explores the contribution of each feature in classifying risk profile (status) using Information Gain (IG) or entropy. Table 4.1 shows the ranking of the features according to their contribution to classifying the risk profiles of the student. The first column (Rank), is the ranking of features from 1 to 24, most significantly contributing (high IG), to the least contributing (lowest IG). The second column is the feature name associated with the ranking. The last column represents the Information Gain (entropy), which is the value $0 \leq e \leq 1$, with 0 as no information gain, and 1 highest IG.

In table 4.1, the features are color-coded differently; biographical characteristics are light blue; pre-college observations are coded light purple; and individual characteristics are blank, is not shaded, as per Tinto [42] framework. The top 3 contributing features are (i) plan description, the student's career choice, (ii) the year started the program, which falls under the individual's characteristics, and (iii) the home language, which is the student's native language. The features ranked from 4 to 7 are (iv) home province, the province/state student originates from; (v) home country, the students country of origin; (vi) the gender of the student; and (vii) the race description; these are biographical characteristics. The top 8 ranked features suggest that biographical characteristics are the most dominant in deducing the student risk status, followed by some few individual attributes. The pre-college attributes show no or minimal effect on student risk profiles.

TABLE 4.1: The Information Gain (entropy) ranking of features.

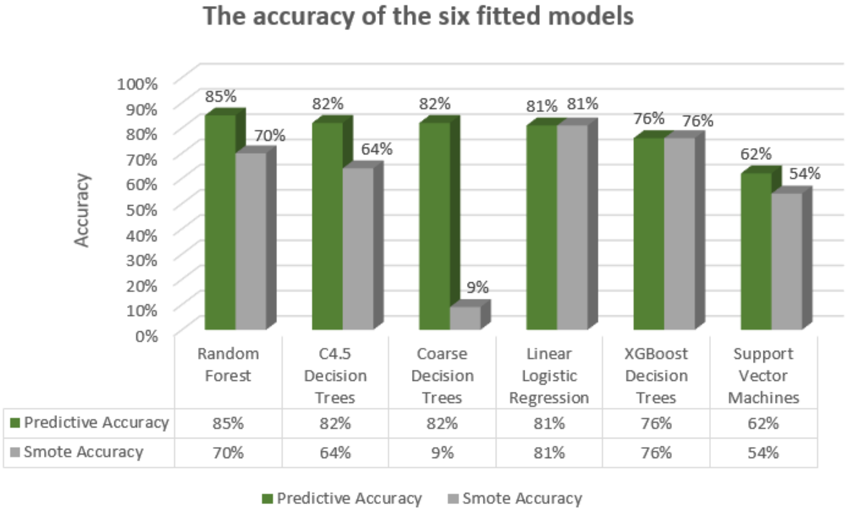
Rank	Feature	Information Gain (e)
1	Plan Description	0.25
2	Year Started	0.24
3	Language	0.12
4	Home Province	0.08
5	Home Country	0.04
6	Gender	0.03
7	Rural or Urban	0.02
8	Race Description	0.02
9	Prob Of Mathematics Streamline	0.00
10	Prob Of Physics Streamline	0.00
11	Prob of Earth Streamline	0.00
12	Prob of Biology Streamline	0.00
13	Aggregate	0.00
14	Number Of Years for Degree	0.00
15	Age at First Year	0.00
16	Quintile	0.00
17	Mathematics Matric Major	0.00
18	English HL	0.00
19	English FAL	0.00
20	Computers	0.00
21	Additional Mathematics	0.00
22	NBTAL	0.00
23	NBTMA	0.00
24	NBTQL	0.00

4.3 Machine Learning Classification Algorithms

In this section, we look at the results from the six fitted machine learning classification algorithms: random forests, linear logistic regression, coarse decision trees, extreme gradient boosted trees, support vector machines, and J48 (C4.5) decision trees. The different metrics that were used to evaluate the predictive performance of our models are confusion matrices, classification accuracy, Kappa statistic, sensitivity, Precision, F-1 score, Area Under the Curve (AUC), and ROC curves.

Classification Accuracy

The classification accuracy is the metric used to measure how many predictions were correctly classified from all the predictions made.



	Random Forest	C4.5 Decision Trees	Coarse Decision Trees	Linear Logistic Regression	XGBoost Decision Trees	Support Vector Machines
🟩 Predictive Accuracy	85%	82%	82%	81%	76%	62%
▨ Smote Accuracy	70%	64%	9%	81%	76%	54%

FIGURE 4.9: The bar graph of the accuracy for the six fitted models.

The accuracy was evaluated using 10-fold cross validation method. Figure 4.9 describes the results of the classification accuracy for the six fitted machine learning classification algorithms: random forests classifier, linear logistic classifier, coarse decision trees classifier, extreme gradient boosted trees classifier, support vector machines classifier, and the J48 (C4.5).

The bars are color coded green represent the predictive accuracy of the models trained with class imbalanced train dataset: lowest risk: 9% , medium risk: 16%, high risk: 67%, highest risk: 8%; and the bars color coded grey represent the smote accuracy, which is the accuracy obtained from the models trained with a corrected class imbalance dataset using SMOTE algorithm, where the minority class is synthetically over sampled, and the majority class is under-sampled, leading to a class balanced dataset: lowest risk: 31%, medium risk: 43%, high risk: 14%, highest risk: 12% . The SMOTE algorithm was introduced (applied) to reduce the class imbalance so that we do not have over estimated training.

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

Figure 4.9 is arranged in descending order by predictive accuracy. Random forest is ranked as number one, which means it has the highest predictive accuracy using the testing dataset. It correctly maps or classifies 85% of the instances from all the predictions made, which means that 425 observations were correctly labeled out of the 500 testing data observations across all classes. When looking at the corresponding smote accuracy, we can see that it achieves 70% classification accuracy, which means than its 15% less accurate compared to the imbalanced class random forest trained model.

Coarse decision trees achieves the second highest predictive accuracy of 82%, following the random forest classifier. This means that the Coarse decision trees correctly deduces 410 instances into the correct (actual) class labels across all the classes. In comparison to the smote accuracy, the Coarse decision trees achieves 9% accuracy, which very low compared to the accuracy of the class imbalanced sets.

The J48 (C4.5) decision trees has the third highest predictive accuracy of 82% after the random forest and the Coarse decision trees. This implies that it can correctly predict 410 instances to the actual risk profiles across all classes. When comparing to the J48 (C4.5) decision trees trained using a smote'd dataset, which means balanced classes, the model achieves 64% accuracy, which is not so bad but yet lower than the accuracy of the class imbalanced set. This could potentially mean that the J48 (C4.5) decision trees model fails to classify risk profiles for balanced datasets correctly, or the synthetic method of generating minority class might affect the dataset and therefore affect the model training.

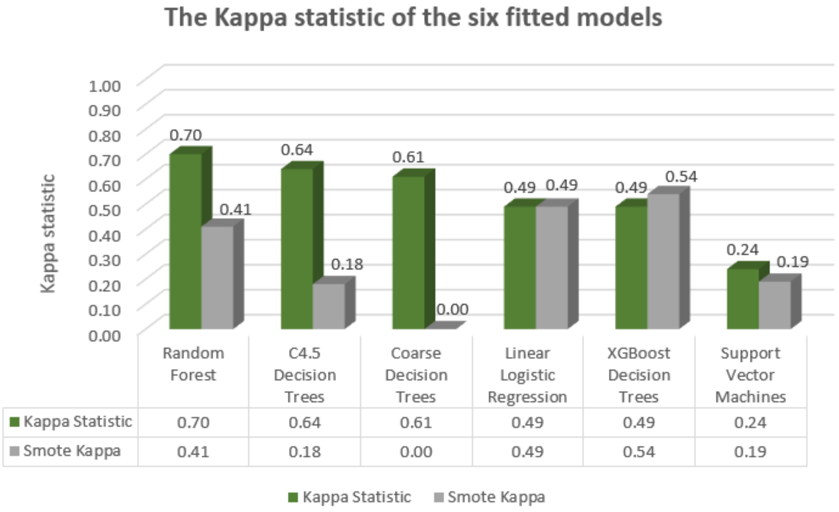
The linear logistic classifier have the fourth highest predictive accuracy of 81% after the random forest, coarse decision trees, and J48 (C4.5) decision trees. The linear logistic classifier correctly maps 405 instances to their actual risk profile classes. When compared to the linear logistic classifier trained with the balanced dataset, smote accuracy is 81%, which is the same as the predictive accuracy achieved with the imbalanced class dataset. This means that the model is not biased in terms of class proportions, whether the data is balanced or not, but it will be able to deduce the risk profile 81% times correctly.

The extreme gradient boosted trees have the fifth highest predictive accuracy of 76% after the random forest, coarse decision trees, J48 (C4.5) decision trees, and the linear logistic classifier. The xgboost correctly maps 380 instances to their actual risk profile classes. When compared to the xgboost trained with the balanced dataset, smote accuracy is 76%, which is the same as the predictive accuracy achieved with the imbalanced class dataset. This means that the model is not biased in terms of class proportions, whether the data is balanced or not, but it will be able to deduce the risk profile 76% times correctly.

The support vector machines have the least predictive accuracy, with predictive accuracy of 62%. This means that the SVM correctly maps 310 instances to their actual risk profile labels. In comparing to the model trained with a balanced dataset, the SVM achieves accuracy of 54%, which is less than the predictive accuracy achieved with an imbalanced dataset. The method used to generate the minority class synthetically could have an impact on the training dataset, leading to low predictive power (accuracy).

Kappa Statistic

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



	Random Forest	C4.5 Decision Trees	Coarse Decision Trees	Linear Logistic Regression	XGBoost Decision Trees	Support Vector Machines
🟩 Kappa Statistic	0.70	0.64	0.61	0.49	0.49	0.24
■ Smote Kappa	0.41	0.18	0.00	0.49	0.54	0.19

FIGURE 4.10: The bar graph of the Kappa statistic for the six fitted models.

Our dataset has imbalanced classes: lowest risk: 9%, medium risk: 16%, high risk: 67%, and highest risk: 8%. The re-sampled dataset using the smote algorithm contains controlled class proportions of lowest risk: 31%, medium risk: 43%, high risk: 14%, and highest risk: 12%.

Figure 4.10 describes the kappa statistics from both the imbalanced dataset and the smote balanced dataset. The bars of the graph are color coded differently, the green bars represents the kappa statistic from models trained with imbalanced datasets, and the grey coded bars represent the kappa statistic from models trained with the class balanced dataset.

The figure 4.10 is ordered by the kappa statistics obtained from the imbalanced dataset models in descending order. The random forest model is ranked number

1, which means it has the highest kappa value of 0.70, which is interpreted as substantial agreement according to [25] interpretation. When comparing the kappa statistic and smote kappa for the same model, random forest, the kappa value for the imbalanced dataset is higher than the one for the balanced class dataset.

The kappa statistic for the coarse decision tree is 0.61, which is interpreted as substantial agreement according to [25] interpretation, which means that this classifier performed substantially across all classes (instances). However, when compared to the smote kappa value of 0.00, it can be seen that for the balanced dataset, the coarse decision tree has poor or no agreement between predicted classification and actual class risk profiles. There is quite a substantial difference between the statistics.

The kappa statistic for the J48 (C4.5) decision tree is 0.64, which is interpreted as substantial agreement according to [25] interpretation, which means that this classifier performed substantially across all classes (instances). However, when compared to the smote kappa value of 0.18, it can be seen that for the balanced dataset, the J48 (C4.5) decision tree has poor or no agreement between predicted classification and actual class risk profiles. There is quite a substantial difference between the statistics.

The linear logistic classifier has a kappa value of 0.49, which is interpreted as moderate agreement according to [25] interpretation, which means that this classifier performed moderately across all classes (instances). The Kappa statistics corresponds (is equal) to the smote kappa statistic, which means that the logistic classifier is not affected by the class imbalance.

The extreme gradient boosted trees model has a kappa statistic of 0.49, which means the agreement is moderate, according to [25] interpretation. This implies that the classifier performed moderately across all risk profile classes. When compared to the smote kappa of 0.54, we find that the kappa for the class imbalanced dataset is lower than that of a smoted dataset.

The support vector machines have a kappa statistic of 0.24, which according to Landis and Koch [25] interpretation, the agreement is fair, meaning the classifier performed fairly across all instances. As it happens with the smote kappa value, which is 0.19, meaning there is a slight agreement across all the classes. The svm model has the least kappa statistic.

Confusion Matrices

In this section, we discuss the results of the confusion matrices of the classification models. A confusion matrix is also known as the error matrix, which is a table used to summarize pre-known predicted labels. It is used to evaluate a classification model using the testing dataset. This metric is used to describe the confusion between the classes. The $n * n$ matrix; rows represent actual class labels, and the columns represent predicted class. The confusion matrices are computed from the models trained with the imbalanced class datasets since the models trained with imbalanced class dataset achieved higher accuracy and higher kappa statistic, which makes them best compared to models trained with balanced class data using smote algorithm. The following tables describe the confusion matrices for the six fitted models and their respective predictive performances.

Table 4.2 illustrates the confusion matrix of the random forest classifier, and the model attains 85% predictive accuracy using the 10-fold cross-validation. This classifier achieves the highest predictive accuracy compared to the other five models fitted on this study. Describing each class performance in detail, we will look at the diagonals that are color coded in green. The diagonals represent the percentage of labels correctly classified into the actual class labels. Beginning from the top cell of the diagonal, we can see that the lowest risk class achieves 30% accuracy; the medium risk class achieves 84% accuracy, high risk class achieves 96%, and highest risk class achieves 55% accuracy. The high risk class achieves the highest predictive accuracy, followed by the medium risk accuracy, then the highest risk class, and the lowest risk class has the lowest predictive accuracy. The lowest risk class and the highest risk class are mostly confused with the medium risk class, with 40% and 27%, respectively. The accuracy and kappa statistic were used to select the best model parameters using the largest values, with mtry = 113, split rule = extra trees, and min.mode.size = 1 as the hyper parameters.

TABLE 4.2: A confusion matrix describing the performance of the random forest model.

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

Table 4.3 illustrates the confusion matrix of the coarse decision tree classifier, and the model achieves 82% predictive accuracy using the 10-fold cross validation. This classifier achieves the second largest predictive accuracy after the random forest, compared to the other fitted models on this study. Looking at the accuracy of each class; the high risk class achieves the highest accuracy of 100%, which means it can correctly distinguish this class from all the other classes; followed by the medium risk class with 68% predictive accuracy; then the highest risk class with 45% accuracy; and the lowest risk class achieves 0% accuracy, meaning that the coarse decision tree model fails to deduce this class. The lowest risk class is mostly confused with the medium risk class. The accuracy and kappa statistic were used to select the best model parameters using the largest values, with cp = 0.05982, split = abs, number of splits = 4, and prune = mr as the hyper parameters.

TABLE 4.3: A confusion matrix describing the performance of the coarse decision tree model.

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

Table 4.4 illustrates the confusion matrix of the J48 (C4.5) decision trees classifier, and the model achieves 82% predictive accuracy using the 10-fold cross validation. This classifier achieves the third highest predictive accuracy after random forest, and coarse decision trees; compared to the other fitted models in this study. Looking at the accuracy of each class; the high risk class achieves the highest accuracy of 96%, which means it can correctly distinguish this class 96% of the time from all the other classes; followed by the highest risk class with 64% predictive accuracy; then the medium risk class with 63% accuracy; and the lowest risk class achieves 20% accuracy. The J48 (C4.5) decision trees model confuses the lowest risk class with the medium and high risk class. The accuracy and kappa statistic were used to select the best model parameters using the largest values, with C = 0.255, and m = 3 as the hyper parameters.

TABLE 4.4: A confusion matrix describing the performance of the J48 (C4.5) model.

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

Table 4.5 illustrates the confusion matrix of the linear logistic classifier, and the model achieves 81% predictive accuracy using the 10-fold cross validation. This classifier achieves the fourth highest predictive accuracy after random forest, coarse decision trees, and J48 (C4.5) decision trees . The diagonals are color coded in green, representing the proportion of labels correctly classified into their correct risk profile classes. The medium risk class and the high risk class are the top performing classes on the logistic regression classifier with 37%, and 88% classification accuracies respectively, followed by the lowest risk profile with an accuracy of 20%, and the highest risk profile class has the least accuracy with 0% correctly classified labels. The logistic regression classifier confuses the lowest risk class and the highest risk class with the medium class and high risk class, as 55% of the highest

risk labels classified as the high risk class. The model fails completely to predict the highest risk class. The accuracy and kappa statistic were used to select the best model parameters using the largest values, with number of iterations = 7 as the hyper parameters.

TABLE 4.5: A confusion matrix describing the performance of the linear logistic regression model.

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

Table 4.6 describes the confusion matrix of the extreme gradient boosted (xgboost) decision trees classifier and the model achieves 76% predictive accuracy using the 10-fold cross validation. This classifier achieves the fifth largest predictive accuracy after random forest, coarse decision trees, J48 (C4.5) decision trees, and the linear logistic classifier compared to the other fitted models on this study. The performance of the xgboost by class shows that the high risk class has the highest accuracy of 95%, followed by medium risk class with 58%, then the lowest risk class has 30%, then the highest risk class is the least performing with an accuracy of 0%. The model fails to recognise this class. The accuracy and kappa statistic were used to select the best model parameters using the largest values, with nrounds = 100, lambda = 1e-04, alpha = 0, and eta = 0.3 as the hyper parameters.

TABLE 4.6: A confusion matrix describing the performance of the XG-Boost tree model.

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

Table 4.7 illustrates the confusion matrix of the support vector machines (SVM) classifier, and the model achieves 62% predictive accuracy using the 10-fold cross validation. This classifier achieves the sixth largest predictive accuracy compared to the other fitted models in this study. The SVM classifier best predicts or classifies the high risk class with 82% accuracy, following with the medium risk class with accuracy 21%. The lowest risk class has 20% accuracy, closing with the highest risk at 18% accuracy. The accuracy and kappa statistic were used to select the best model parameters using the largest values, with kernel = linear, and cost = 0.5 as the hyper parameters.

TABLE 4.7: A confusion matrix describing the performance of the support vector machine model.

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

Sensitivity/Recall Metric

The sensitivity or recall is the measure of the proportion of actual positives that are correctly classified. Table 4.8 illustrates the recall metric for the six trained models and each risk profile class. This will help us describe which models correctly classify risk profiles and which risk profile classes have a higher proportion of correctly classified labels.

The model that has the best (highest) overall recall is the random forest (0.66), followed by the J48 (C4.5) decision trees (0.61), then the coarse decision trees (0.53), then the linear logistic classifier (0.51), then the xgboost decision tree (0.46), and the SVM has the least recall (0.35). This implies that the random forest has the highest proportion of correctly classified risk profile labels. The svm the least recall, meaning that they have a lower proportion of classes correctly classified as their actual label.

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

TABLE 4.8: The Sensitivity (Recall) of the six trained models.

Model	**Lowest**	**Medium**	**High**	**Highest**
Random Forest	0.30	0.84	0.96	0.55
Linear Logistic Regression	0.40	0.70	0.95	0.00
Coarse Decision Trees	0.00	0.68	1.00	0.45
Extreme Gradient Boosted Trees	0.30	0.58	0.95	0.00
Support Vector Machines	0.20	0.20	0.82	0.18
J48 (C4.5) Decision Trees	0.20	0.63	0.96	0.64

Precision Metric

The precision refers to the percentage of the results which are relevant, meaning that if the model predicts a true class, how often is it correct? Table 4.9 illustrates the results of the precision metric for the six fitted models, and the different class levels (risk profiles). The higher the precision value, the better the model is at predicting relevant risk profiles quite often. In table 4.9, the first column is the model name or description; and the following four columns represent the risk profile classes: lowest risk, medium risk, high risk, and highest risk profile.

TABLE 4.9: The Precision of the six trained models.

Model	Lowest	Medium	High	Highest
Random Forest	0.60	0.62	0.93	1.00
Linear Logistic Regression	0.40	0.47	0.90	NA
Coarse Decision Tree	NA	0.50	0.90	1.00
Extreme Gradient Boosted Tree	0.38	0.50	0.89	0.00
Support Vector Machines	0.13	0.33	0.80	0.18
J48 (C4.5) Decision Trees	0.33	0.55	0.95	0.64

We will first look at the performance of each model for each risk profile class. The random forest model has the highest overall precision (0.79), and the highest across all risk profile classes; followed by the J48 (C4.5) decision trees (0.62), then the coarse decision trees (0.60), then the linear logistic classifier (0.44), then the xgboost (0.44), and the least is the svm (0.36). The svm has the least precision meaning that this models prediction's are often not correct, high miss classification rate.

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

F-Score

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

Table 4.10 is the F score of the six fitted classification models. The first column is the model name or description, and the following four columns represent the risk profile classes: lowest risk, medium risk, high risk, and highest risk profile.

TABLE 4.10: The F score of the six trained models.

Model	**Lowest**	**Medium**	**High**	**Highest**
Random Forest	0.40	0.71	0.95	0.71
Linear Logistic Regression	0.40	0.56	0.92	NA
Coarse Decision Tree	NA	0.58	0.95	0.63
Extreme Gradient Boosted Tree	0.33	0.54	0.92	NA
Support Vector Machines	0.16	0.26	0.81	0.18
J48 (C4.5) Decision Trees	0.25	0.59	0.96	0.64

In the table 4.10, random forest has the highest F score when compared to the other five fitted models; and the svm has the lowest F score, which means on overall, the model is not performing well in classifying the risk profile classes.

Area Under the Curve

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

Table 4.11 describes the results of the AUC for each of the six trained models. The first column is the model name, and the second column is the AUC value in descending order. The top-ranked model has the highest AUC when compared to the other models, which is the random forest with AUC of 0.95; followed by the xgboost decision tree with AUC of 0.92; then the coarse decision tree with AUC of 0.91; then the J48 (C4.5) decision trees with AUC of 0.91; then the linear logistic classifier with AUC of 0.83; and the model with the least AUC is the SVM with AUC of 0.77.

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

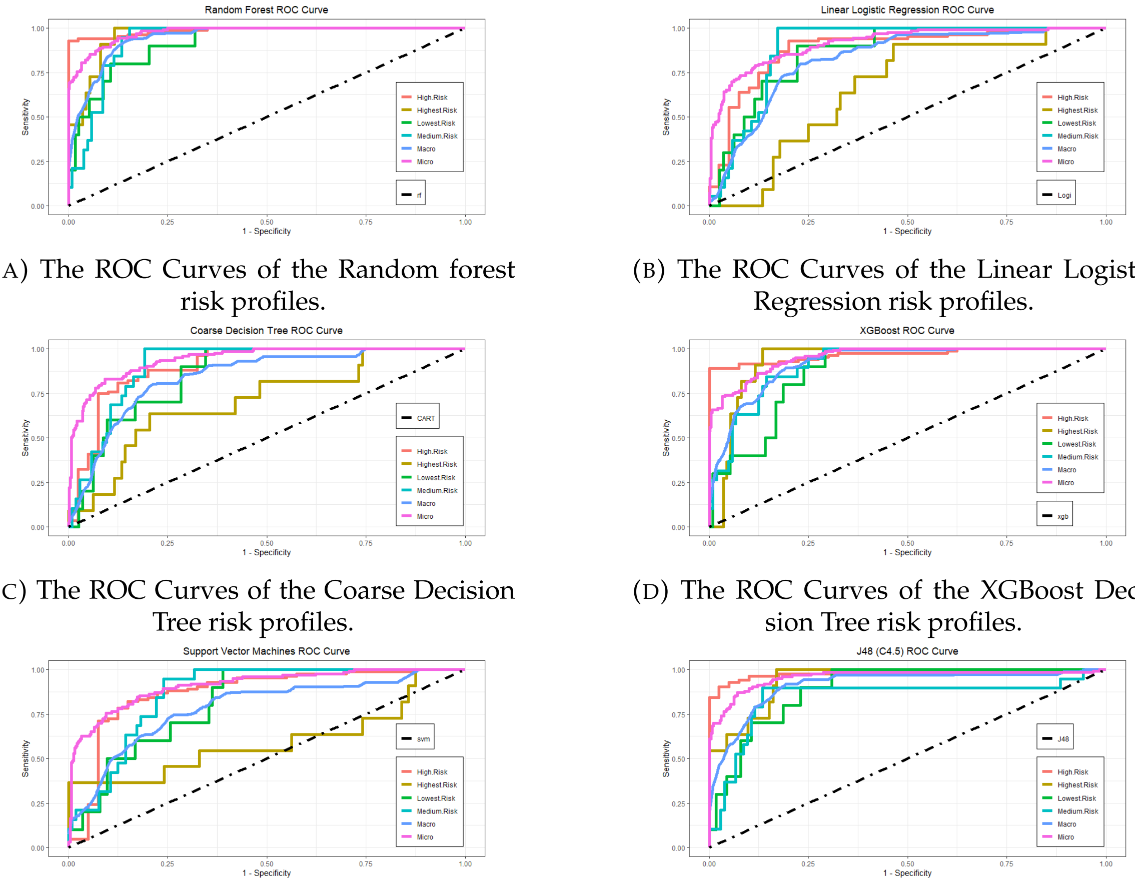
TABLE 4.11: The Area Under the Curve (AUC) of the six fitted models.

Model	**AUC**
Random Forest	0.95
Extreme Gradient Boosted Tree	0.92
Coarse Decision Tree	0.91
J48 (C4.5) Decision Trees	0.91
Linear Logistic Regression	0.83
Support Vector Machines	0.77

Receiver Operating Characteristics (ROC) Curve

The Area Under the Curve (AUC) – Receiver Operating Characteristics (ROC) curve is a function of sensitivity plotted against 1- specificity at a specific decision threshold. The main aim of the AUC-ROC curve is to maximize the sensitivity and 1 - specificity.

The ROC curve represents the probability or score curve. The AUC is the measure of separability between classes. The higher the AUC, the better the model is at predicting class 1 as class 1 and class 2 as class 2. For the multi-class problem with N classes, we have N ROC curves, representing N probability curves.



(A) The ROC Curves of the Random forest risk profiles.

(B) The ROC Curves of the Linear Logistic Regression risk profiles.

(C) The ROC Curves of the Coarse Decision Tree risk profiles.

(D) The ROC Curves of the XGBoost Decision Tree risk profiles.

(E) The ROC Curves of the Support Vector Machines risk profiles.

(F) The ROC Curves of the J48 (C4.5) Decision Tree risk profiles.

FIGURE 4.11: The ROC curves of the six fitted models.

Figure 4.11 illustrates the ROC curves of each class for all fitted machine learning classification models; also including the micro class which was created by combining all the classes together, thereby converting the multiclass classification into binary classification problem; and also adding the macro class which was calculated by taking the average of all the class results, thereby converting to one versus the rest and then applying linear interpolation between points of the ROC curve. This then totals six roc curves for the six classes that are color-coded differently; the lowest risk is coded green, the medium risk is blue, high risk is orange, highest is yellow, the micro class is pink, and macro class is powder blue.

The random forest roc curve in figure 4.11a, describes the six roc curves for the four risk profile classes, micro class, and the macro class. The high risk class is the best performing class when compared to the other classes. In figure 4.11b, we describe the performance of the different risk profile classes of the logistic classifier, the medium risk class outperforms the other classes. The micro class is the best performing class in the coarse decision tree, figure 4.11c, which means that stacking all the classes together yields better results in the coarse decision tree. The extreme gradient boosted trees roc curve, figure 4.11d shows that the high risk class better performs than the other classes. The micro class is the best performing class for both the support vector machines and the J48 (C4.5) decision trees, implying that combining the four risk profiles yields better results for both these models. The rest of the J48 (C4.5) decision trees classes (figure 4.11f) lie on the margin (benchmark), which means that the classification to those classes is random, or the model randomly allocates the predicted labels to these classes: lowest risk, medium risk, high risk, highest risk, and the synthetic class macro class.

4.4 Discussion

The previous sections discussed the main results obtained from this study. This section will look at the contribution of this research in the field of education and computational sciences.

The main aim of this research is to explore the relationship between background, individual, and pre-college characteristics on learner attrition, as per the model of leaner attrition by Tinto [42]. These characteristics are then used as input attributes to predict the student attrition by classifying the students into four risk profiles: 'Lowest risk' - where the student is expected to complete their degree in the minimum time (3 years); 'medium risk' - where the student is expected to complete in more than the minimum time; 'high risk' - where the student fails or drops-out before the minimum time; and 'highest risk' - where the student fails in more than the minimum time.

Our study of deducing the students into the correct risk profiles using biographical (background), individual, and schooling characteristics; showed that student attrition or classifying the students into the right risk profiles is dominantly affected by biographical characteristics; followed by individual attributes. The pre-college characteristics show minimal or no effect on deducing student risk profiles. Similar results were achieved by Ajoodha and Jadhav [4], that the eight most significant (contributing) attributes, are biographical and individual characteristics play a major (important) role in deducing the student into the correct risk profiles, as per Tinto [42] conceptual model. Moreover, studies by [Abu Tair and El-Halees [3], Ajoodha and Jadhav [4], Pal [36], Pandey and Pal [37], Downs et al. [17], Mwamwenda [32], Steenkamp, Baard, and Frick [40], Alfan and Othman [6], Yukselturk, Ozekes, and Türel [49], and Pandey and Pal [37]] also attest to these results, that family or background characteristics have a high influence on student attrition. The impact of using a subset of features (i.e., features with higher information gain), in terms of classification accuracies and other measures, would still need to be investigated.

The observations from the results (section 4.3), in general, demonstrate that the fitted models perform well on an imbalanced class dataset; compared to models fitted with controlled balanced class dataset using SMOTE algorithm, where minority class is synthetically over sampled and majority class is under sampled. This can be due to the process applied by the smote algorithm while generating synthetic samples, smote does not take into consideration the neighbouring samples from other classes. It can then result in overlapping of classes and can introduce additional noise.

The fitted machine learning algorithms were successfully able to detect (deduce) the different risk profiles. However, the positive class detection rate differs for different class proportions, i.e., majority class (67%), high risk profile has higher rates of positive (correct) detection of this class; compared to the minority classes (9% and 8%), lowest and highest risk profiles respectively, have a high negative rate (miss classification); across all the models. The impact of skewed class sizes of the training and testing set would need to be investigated on the classification accuracy.

The random forest model achieved the best results with an accuracy of 85%, kappa statistic of 0.7, overall precision of 0.79, overall recall of 0.66, F-score of 0.69, and an AUC of 0.95 over the four risk profiles. The accuracy and kappa statistic were used to select the best model parameters using the largest values, with mtry = 113, split rule = extra trees, and min.mode.size = 1 as the hyper parameters.

The poorest of the results were achieved by the support vector machines, with accuracy 62%, kappa statistic of 0.24, overall precision of 0.36, overall recall of 0.35, F-score of 0.35, and an AUC of 0.77 over the four risk profiles. The hyper parameters used to select the best model with highest accuracy and kappa statistic, with kernel = linear, and cost = 0.5 as the hyper parameters.

The results are in line with what other authors discovered; [4, 34] achieved accuracies ranging from 69% to 76% using background, individual, and schooling factors to predict student attrition applying data mining algorithms like Naive Bayes, decision trees, random forests, linear logistic regression, and support vector machines.

The random forest outperformed the other models because its algorithm creates multiple decision trees and merges them together to obtain a more stable and accurate prediction. This then avoids over fitting of trees in the model. The support vector machine may have poorly performed because it works well with linearly separable datasets, and linearity is no practically applicable on most real datasets.

4.5 Summary

In this chapter 4, we have discussed the results obtained from applying our methods discussed in chapter 3. The results describe the output from our trained models. We found essential or most significant features in classifying risk profile status. The top eight important features provide evidence that biographical characteristics are the most dominant in deducing the student risk profile. Models trained with imbalanced class datasets proved to be better performing in terms of accuracy, kappa statistic, ability to separate different classes when compared to models trained with balanced class dataset using Synthetic Minority Over-sampling Technique (smote) Algorithm. The random forest model outperforms the other models with an accuracy of 85%, kappa of 0.7, AUC of 0.95, and the roc curves closer to the top left corner; when compared to the linear logistic classifier, xgboost decision tree, SVM, coarse decision tree, and the J48 decision tree. The following chapter 5 will look at the conclusions and future works.

Chapter 5

Conclusions and Future Work

This chapter provides an overview of the research, the summary of the results and discussions put forward, and finally the future work, where the research can be improved and done differently to achieve better results.

5.1 Conclusions

This study presents a discussion about predicting student attrition using background, individual, and schooling attributes as per Tinto et al. [43] framework. A synthetic dataset simulated using the Bayesian Network of student's background, individual, and schooling characteristics were used as input features to predict student risk profile. Matlab and R studio are used as the machine learning tools and graphics interface (visualization) tools. Various machine learning algorithms are fitted (trained): random forest, coarse decision trees, linear logistic regression, J48 (C4.5) decision trees, extreme gradient boosting trees, and support vector machines.

The dataset has an imbalanced class set, which led to two training sets, whereby one set is a standard imbalanced class dataset, and the other is a balanced class set using the SMOTE algorithm. The models were then trained using both sets of datasets. The models that were trained with an imbalanced class set performed far better than models that were fitted using a balanced class dataset. The random forest model performed best compared to the other fitted models, with an accuracy of 85%, kappa statistic of 0.70, higher recall, precision, and F-score across all the risk profile classes; and an area under the curve (AUC) of 0.95.

Information gain (IG) shows that predicting student attrition or rather deducing the student into the correct risk profile is dominantly affected by biographical characteristics, followed by individual attributes, and the pre-college (schooling) characteristics show minimal or no effect on student attrition.

This research contributes to an argument for a more complex view of predicting undergraduate student attrition by including the student's biographical, individual, and schooling characteristics. We reject the application of only pre-college attributes (i.e., matric results and or National Bench Mark tests) as a primary schooling system measure for student placement at university or higher education institutions.

The study concludes that student attrition is affected by biographical and individual attributes, and therefore these factors should be taken into consideration in the higher education placement system.

5.2 Future Work

The study of student attrition is one of the most important studies in the educational sciences. In the future, we would like to work with actual real student enrollment data. This will help us verify if our theoretical model is applicable in real-world data, or can distinguish the different risk profiles. It will also shed light on the most important features or characteristics that affect student attrition.

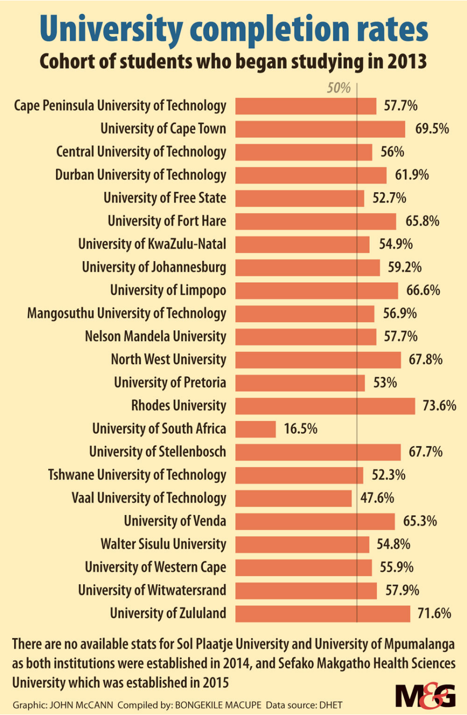
Currently, our method used the duration it took to complete the degree (qualification), to derive the target variable, risk profile; for example, the longer it takes to finish the degree, the higher the risk profile. The future study will use the target variable attrition (dropout) or not, and apply binary classification modeling, then use the model to get predicted probabilities of attrition; then calibrate (bin) these probabilities of attrition; for example: $0\% - 25\%$ - bin 1, and is the lowest risk profile; $26\% - 50\%$ - bin 2, medium risk profile; $51\% - 75\%$ - bin 3, high risk profile; and $76\% - 100\%$ - bin 4, highest risk profile. Then evaluate how well the model is able to deduce students into the correct risk profiles.

We will also build an application where we will deploy our model, such that, when a student inputs all their background, individual, and schooling attributes, the model will output a prediction indicating the risk profile of the student.

Appendix A

The Supporting Figures and Tables

A.1 The Current Attrition Rates



There are no available stats for Sol Plaatje University and University of Mpumalanga as both institutions were established in 2014, and Sefako Makgatho Health Sciences University which was established in 2015

Graphic: JOHN McCANN Compiled by: BONGEKILE MACUPE Data source: DHET

FIGURE A.1: The university completion rates on a cohort of students who enrolled for first time studies in 2013 [27].

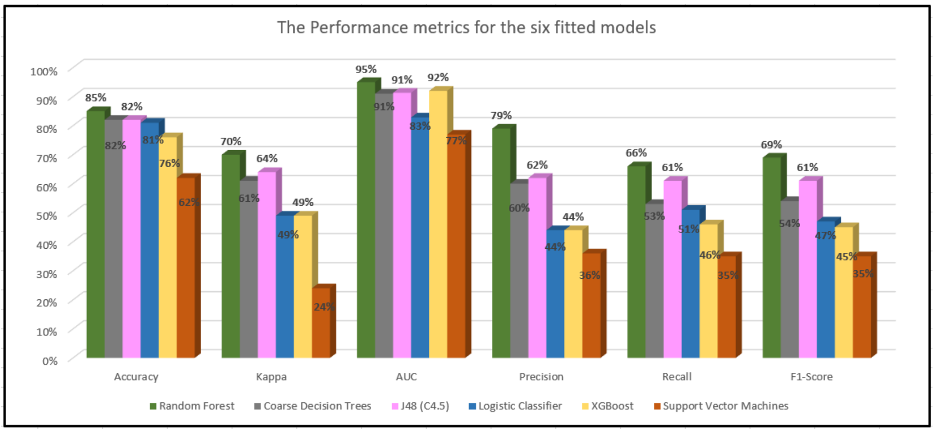


FIGURE A.2: The bar graph of summary of all the performance metrics.

A.2 The Description of Features

TABLE A.1: The description of the features and their possible values.

Attribute	Description	Possible Values
Risk Status	The risk profile (target variable)	{low, medium, high, highest}
Gender	sex identity	{Male, Female}
Race	Race description	{Black, White, Coloured, Indian, Chinese}
Age at first year	Age of the student at first year	{15 to 60}
Home Language	Native language	{ All national official languages}
Home Province	The province (state) the student originates	{The 9 south african provinces, and the other national states }
Home Country	The originating country	{All national countries}

Continued on next page

Table A.1 – continued from previous page

Attribute	**Description**	**Possible Values**
Year Started	The year started degree	{ 2008 to 2018}
Plan Description	The career choice	{ All science career choices.}
Aggregate	The aggregate of marks	{ 0 to 100 }
Years in degree	Number of years in the degree	{ 1 to 13}
Prob of streamline (maths, physics, earth, biological)	The probability of being successful at a particular streamline	{ 0 to 1 }
Rural or urban	The school location	{ urban, rural }
School quintile	The poverty ranking of schools (quintile)	{1 to 5}
Mathematics major	The core mathematics	{ 0 to 100 }
Additional Mathematics	The technical mathematics	{ 0 to 100 }
English FAL	English first additional language	{ 0 to 100 }
English HL	English home language	{ 0 to 100 }
Computers	Computer subject	{ 0 to 100 }
NBTAL, NBTMA, NBTQC	National benchmark tests	{ 0 to 100 }

Bibliography

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

[2] R Abilash. *Applying random forest (classification) - Machine learning algorithm from scratch with real datasets*. https://medium.com/\spacefactor\@m{}ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c57. Accessed: 2019-10-02.

[3] Mohammed M Abu Tair and Alaa M El-Halees. "Mining educational data to improve students' performance: a case study". In: *Mining educational data to improve students' performance: a case study* 2.2 (2012).

[4] Ritesh Ajoodha and Ashwini Jadhav. "Identifying at-risk undergraduate students using Biographical and Enrollment Observations for Mathematical Science Degrees at a South African University". In: *Arctic Journal* 72.7 (2019), pp. 42–71. ISSN: 0004-0843.

[5] Ritesh Ajoodha, Ashwini Jadhav, and Shalini Dukhan. "Forecasting Learner Attrition for Student Success at a South African University". In: *In Conference of the South African Institute of Computer Scientists and Information Technologists 2020 (SAICSIT '20), September 14-16, 2020, Cape Town, South Africa. ACM, New York, NY, USA, 10 pages.* ACM. 2020. DOI: https://doi.org/10.1145/3410886.3410973.

[6] Ervina Alfan and Nor Othman. "Undergraduate students' performance: the case of University of Malaya". In: *Quality assurance in education* 13.4 (2005), pp. 329–343.

[7] Ervina Alfan and Nor Othman. "Undergraduate students' performance: the case of University of Malaya". In: *Quality Assurance in Education* 13.4 (2005), pp. 329–343. DOI: 10.1108/09684880510626593. eprint: https://doi.org/10.1108/09684880510626593. URL: https://doi.org/10.1108/09684880510626593.

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

[9] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[10] Brijesh Kumar Bhardwaj and Saurabh Pal. "Data Mining: A prediction for performance improvement using classification". In: *arXiv preprint arXiv:1201.3418* (2012).

[11] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[12] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[13] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794.

[14] Tianqi Chen et al. "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* (2015), pp. 1–4.

[15] Line Clemmensen et al. "Sparse discriminant analysis". In: *Technometrics* 53.4 (2011), pp. 406–413.

[16] Evgenia Dimitriadou et al. "Package 'e1071'". In: *R Software package, avaliable at http://cran. rproject. org/web/packages/e1071/index. html* (2009).

[17] Colleen Thelma Downs et al. "Investigation into the academic performance of students in Bioscience at the University of Natal, Pietermaritzburg, with a particular reference to the Science Foundation Programme students". In: (2002).

[18] Giuliano Galimberti, Gabriele Soffritti, Matteo Di Maso, et al. "Classification trees for ordinal responses in R: the rpartScore package". In: *Journal of Statistical Software* 47.i10 (2012).

[19] Syed Tahir Hijazi and SMM Naqvi. "Factors affecting students' performance." In: *Bangladesh e-journal of sociology* 3.1 (2006).

[20] Kurt Hornik et al. "The rweka package". In: (2007).

[21] Reid A Johnson et al. "A data-driven framework for identifying high school students at risk of not graduating on time". In: *Bloomberg Data for Good Exchange Conf*. Vol. 5. 2015.

[22] Dorina Kabakchieva. "Predicting student performance by using data mining methods for classification". In: *Cybernetics and information technologies* 13.1 (2013), pp. 61–72.

[23] Max Kuhn. "The caret package". In: *R Foundation for Statistical Computing, Vienna, Austria. URL https://cran. r-project. org/package= caret* (2012).

[24] Jakub Kuzilek et al. "OU Analyse: analysing at-risk students at The Open University". In: *Learning Analytics Review* (2015), pp. 1–16.

[25] J Richard Landis and Gary G Koch. "The measurement of observer agreement for categorical data". In: *biometrics* (1977), pp. 159–174.

[26] Elaine Lust and Frances C Moore. "Emotional intelligence instruction in a pharmacy communications course". In: *American journal of pharmaceutical education* 70.1 (2006), p. 06.

[27] Bongekile Macupe. *Report finds most SA universities well run*. https://mg.co.za/article/2019-11-15-00-report-finds-most-sa-universities-well-run. Accessed: 2019-11-25. 2019.

[28] C Manliguez. "Generalized confusion matrix for multiple classes". In: (2016).

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

[30] Dell Michael and Dell Susan. *The Economics of Completion*. 2017. URL: {\urlhttps://impact.msdf.org/university-completion/}. accessed: 05.08.2019.

[31] Sharma Mohna. *Support Vector Machines – Not for the faint-hearted*. URL: \urlhttps://storybydata.com/datacated-weekly/support-vector-machines-not-for-the-faint-hearted/.

[32] Tuntufye S Mwamwenda. "Gender differences in scores on test anxiety and academic achievement among South African University graduate students". In: *South African Journal of Psychology* 24.4 (1994), pp. 228–230.

[33] Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy. "A comparative analysis of techniques for predicting academic performance". In: *2007 37th annual frontiers in education conference-global engineering: knowledge without borders, opportunities without passports*. IEEE. 2007, T2G–7.

[34] Edin Osmanbegovic and Mirza Suljic. "Data mining approach for predicting student performance". In: *Economic Review: Journal of Economics and Business* 10.1 (2012), pp. 3–12.

[35] Mahesh Pal. "Random forest classifier for remote sensing classification". In: *International Journal of Remote Sensing* 26.1 (2005), pp. 217–222.

[36] Saurabh Pal. "Mining educational data using classification to decrease dropout rate of students". In: *arXiv preprint arXiv:1206.3078* (2012).

[37] Umesh Kumar Pandey and Saurabh Pal. "Data Mining: A prediction of performer or underperformer using classification". In: *arXiv preprint arXiv:1104.4163* (2011).

[38] Vili Podgorelec et al. "Decision trees: an overview and their use in medicine". In: *Journal of medical systems* 26.5 (2002), pp. 445–463.

[39] J Ross Quinlan. "Combining instance-based and model-based learning". In: *Proceedings of the tenth international conference on machine learning*. 1993, pp. 236–243.

[40] LP Steenkamp, RS Baard, and BL Frick. "Factors influencing success in first-year accounting at a South African university: A comparison between lecturers' assumptions and students' perceptions". In: *South African Journal of Accounting Research* 23.1 (2009), pp. 113–140.

[41] Endya B Stewart. “School structural characteristics, student effort, peer associations, and parental involvement: The influence of school-and individual-level factors on academic achievement”. In: *Education and urban society* 40.2 (2008), pp. 179–204.

[42] Vincent Tinto. “Dropout from higher education: A theoretical synthesis of recent research”. In: *Review of educational research* 45.1 (1975), pp. 89–125.

[43] Vincent Tinto et al. “Building community.” In: *Liberal Education* 79.4 (1993), pp. 16–21.

[44] L Torlay et al. “Machine learning–XGBoost analysis of language networks to classify patients with epilepsy”. In: *Brain informatics* 4.3 (2017), p. 159.

[45] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. “The max-min hill-climbing Bayesian network structure learning algorithm”. In: *Machine learning* 65.1 (2006), pp. 31–78.

[46] Jarek Tuszynski and Maintainer ORPHANED. *Package ‘caTools’*. 2013.

[47] Laura P Womble. “Impact of stress factors on college students academic performance”. In: *Undergraduate journal of Psychology* 16.1 (2003), pp. 16–23.

[48] Marvin N Wright and Andreas Ziegler. “ranger: A fast implementation of random forests for high dimensional data in C++ and R”. In: *arXiv preprint arXiv:1508.04409* (2015).

[49] Erman Yukselturk, Serhat Ozekes, and Yalın Kılıç Türel. “Predicting dropout student: an application of data mining methods in an online education program”. In: *European Journal of Open, Distance and e-learning* 17.1 (2014), pp. 118–133.

[50] Mark H Zweig and Gregory Campbell. “Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.” In: *Clinical chemistry* 39.4 (1993), pp. 561–577.