

https://dspace.org/

A theoretical model to predict undergraduate attrition based on background and enrollment characteristics

Mathye, Macdaline Raisibe

2020

https://hdl.handle.net/10539/30996

Downloaded from a DSpace repository. https://dspace.org/

A Theoretical Model To Predict Undergraduate Attrition Based On Background And Enrollment Characteristics

---

Macdaline Raisibe Mathye
1887635

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

5 November 2020

Declaration

I, Macdaline Raisibe Mathye with student number 1887635 declare that:

The Research Report submitted for the degree of Master of Science in the field of e-Science at the University of the Witwatersrand, Johannesburg for this academic year is to the best of my knowledge, my individual unaided work.

It has not been submitted for any degree or examination at any other university. Where, discussion has been informed by previously-submitted work, this has been indicated as such.



Macdaline Raisibe Mathye
1887635
5 November 2020

Dedication

For the degree of Master of Science is dedicated to my family, my late mother Agnes Mavhungu Mathye, my sisters, my siblings, as well as my father William Mathye for their valuable time, support and assistance throughout. My extended open heart dedication is to My Lord Jesus Christ.

Acknowledgements

As in mainly, my gratitude extends to the Lord, the creator of the universe.

My respect and acknowledgment to several mentioned helpers below, because the research report's success was possible all due to the great contribution and guidance given to me, the researcher:

University of the Witwatersrand, for giving the researcher the enormous opportunity and space, in order to conduct research at the institution. I am honourned for the opportunity.

The **DST-CSIR National e-Science Postgraduate Teaching and Training Platform (NEPTTP)**, for the award of receiving funding for both the academic years of 2018-2019. I as the researcher, i am truly grateful for the opportunity, without your funding none of this research would have been possible.

Dr. Ritesh Ajoodha, the supervisor, his appreciate time, guidance, patience, support and assistance throughout the study made the study to be successful. Under his convenient supervision, the researcher was able to make enormous progress.

Dr. Ashwini Jadhav, the co-supervisor, also due to her valuable input and time, support and assistance throughout the research, the study was a success.

Abstract

Developing graduate readiness amongst students who enters university with risk factors is one of the greatest challenges of institutions. Evidence that students with risk profiles are not likely to seek assistance when required complicates the problem. In this work we aim to identify the profiles of students with attributes indicating learner vulnerability. A synthetic higher education dataset from 2008-2018 was used for the purpose of this research. We follow the conceptual framework by Tinto (1975) to deduce student attrition.

The features considered were academic courses, grade 12 marks, background information, individual attributes and respective outcomes for science students. To identify profiles of vulnerable students, several machine learning classification models to deduce the learner into four risk classes: Lowest Risk, Medium risk, High risk and Highest risk were used. The analysis used various predictive models: Random Forests, Decision trees, Support vector Machines, Bayesian Network classifier and multinomial Logistics regression. Effectiveness of each model was tested through 10-Fold Cross Validation and all the hyperparameters were tuned. The Random Forest performed the best with an accuracy of 73% and the least predictive model with 63% was the Multinomial Logistic Regression. The major contribution of this report are: a) a comparison of predictive models to calculate the probability of a learner's risk profile, by contextualizing the students synthetic background, individual and schooling data. b) a ranking of employed features according to their entropy to correctly classify the class variable.

keywords: Learner vulnerability, Attrition, Background characteristics; Individual attributes; Pre-college or schooling attributes; Machine learning models

Contents

Declaration i

Dedication ii

Acknowledgements iii

Abstract iv

List of Figures vii

List of Tables viii

1 Introduction 1
- 1.1 Problem Statement . . . 1
  - 1.1.1 Research Questions . . . 1
  - 1.1.2 Objectives . . . 2
- 1.2 Research Purpose . . . 2
- 1.3 Literature . . . 2
- 1.4 Methodology . . . 3
- 1.5 Contributions . . . 3
- 1.6 Overview . . . 3

2 Literature Review 4
- 2.1 Features . . . 7
- 2.2 Model and Accuracies . . . 8

3 Research Methodology 10
- 3.1 Data . . . 11
- 3.2 Features . . . 13
- 3.3 Descriptive Statistics . . . 14

3.4 Research Models . . . 14
3.4.1 Random Forests . . . 15
3.4.2 Decision Trees . . . 15
3.4.3 Support Vector Machines . . . 15
3.4.4 Multinomial Logistic . . . 16
3.4.5 Bayesian Network Classifier . . . 16
3.5 Methods . . . 17
3.6 Analysis and Evaluation . . . 17
3.6.1 Metrics . . . 18
3.6.2 Comparisons . . . 19
3.6.3 Training/Testing Split . . . 19
3.6.4 Feature Selection . . . 19
3.7 Limitations . . . 20

4 Results and Discussion **21**
4.1 Analysis of data . . . 21
4.1.1 Descriptive statistics . . . 22
4.1.2 Feature Information Gain . . . 26
4.1.3 Results of models: Confusion Matrix and Accuracies . . . 28
Confusion Matrices: . . . 28
Model Accuracies, Recall and Precision . . . 31
Recall . . . 32
Precision . . . 33
F1-score . . . 34
4.1.4 Discussion . . . 34
4.1.5 Summary . . . 35

5 Conclusions and Future Work **36**
5.1 Conclusion . . . 36
5.2 Future work . . . 38

Bibliography **39**

List of Figures

2.1 Conceptual framework model of Tinto [28] showing the relationship between background characteristics, individual attributes, and pre-schooling attributes to the drop out decisions. 5

3.1 This Bayesian Network Structure to predict learner vulnerability . . 12

4.1 The pie chart of the gender and race description of the learner. 22
(a) Gender 22
(b) Race Description 22
4.2 The distribution showing the relationship between Risk Status and different variables: Quintile of the learners, Race Description, Gender and first year outcomes 23
4.3 The relationship between Aggregate, Race Description and the Risk Status. 24
4.4 Total count of each response variable (Risk Status). The response variable represents: Lowest Risk, Medium Risk, High Risk and Highest Risk. 25

List of Tables

2.1 A table relating the fundamental factors to review of key authors who used varying feature sets and models to predict learner attrition 6

3.1 A list of attributes 14

4.1 A ranking of the information gain (entropy, denoted *e*) for a set of features to predict learners Risk Profile. The top 11 features are highlighted in light blue. 27
4.2 The predictive accuracy of the six trained models. 31
4.3 The Recall of the predictive Models 32
4.4 The Precision of the predictive Models 33
4.5 The F1-score of the predictive Models 34

Chapter 1

Introduction

This chapter introduces the study of theoretically predicting undergraduate attrition based on background and enrollment characteristics. The problem statement, research purpose and questions, overview of the methods and contributions of the study will be looked into in this chapter.

1.1 Problem Statement

Attrition, dismissal, termination of courses is widespread in higher education institutions [23]. Because of an increasing number of vulnerable students who enters university, developing graduate-ready students is one of the greatest challenges of institutions [10]. Transiting into tertiary environment and its education system is a struggle to first year students because they find themselves under-privileged of the essential skills required in their field and a cultural capital for the pursuit of their studies. Due to the incompatibility with the chosen curriculum most students who are admitted into university programmes fail to complete their degrees [2].

1.1.1 Research Questions

In this report we attempt to answer the following research questions:
Can we deduce learner attrition by using certain characteristics? Which of the adopted classification models are most suitable for classifying learners into risk profiles using these potential factors?

1.1.2 Objectives

The objectives of this study are: a) use background, individual, and schooling characteristics to predict risk profiles; b) train or build predictive classification models using background, individual, and schooling characteristics; c) compare results of the models with previous literature results.

1.2 Research Purpose

Students with vulnerability are considered to have higher probability of not progressing well in their studies, that is, dropping out of university or failing academically [10]. The issue is complicated by the evidence showing that learners who are vulnerable are less likely to seek support when they need it [21]. There is a growing concern with attrition and low throughput rates in universities [2]. Surviving the hardships of academic life is difficult for students as they lack the necessary skills and background even though the students starts as freshman having the potential to succeed but end up as being vulnerable [21]. To empower them, an effective method must be implemented where they can be provided with proper assistance when requested. In this research we attempt to predict learner attrition to identify student vulnerability by using background, individual and schooling characteristics so we can implement or provide proper interventions that is meaningful and as cost effective as possible. Providing these interventions to the learners, we can alleviate the possibility that the learner will fail their selected programmes. [2].

1.3 Literature

Education research has been investigated often recently [8, 27]. A study done by [8] on attrition included part-time attendance type, mature age and non-English speaking background on undergraduate data as features with models stepwise multiple regression analysis, analysis of variance and logistic regression. While [2] attempted to use background, individual, and schooling learner features to deduce student attrition at a South African institution using data from a South African Institution, the authors used several machine learning classification models with confusion matrices to gauge model performance. The effectiveness of being unable to

learn due to lack of foundation closes significant economic, academic and social opportunities to the students [10].

1.4 Methodology

In this research report we use several machine learning predictive models and evaluation metrics. The study uses background, individual and schooling characteristics to predict the distribution over risk profiles (response variable) as similar to the Tinto framework [28]. The response variable has different Risk Profile namely: Lowest Risk, Medium Risk, High Risk, and Highest Risk. We use the following predictive machine learning models: Random Forests, Decision trees, Support vector Machines, Bayesian Naive classifier and Multinomial Logistics regression. To evaluate the predictive performance of our models we use confusion matrices, classification accuracy, precision, recall and F1-score.

1.5 Contributions

This research contributes to the current body of knowledge by:

- Providing a predictive model that will be able to predict learner vulnerability.
- A comparison of features based on the entropy to predict the class variables.

1.6 Overview

In the remaining chapters or sections of this research, we consider or focus on the following: Chapter 2 which reviews the work that has been done in the field and their results; Chapter 3 highlights on the data and research methodology to achieve the set research aim; Chapter 4 shows the findings and discuss on it. Chapter 5 concludes the research and puts forward recommendations of future work.

Chapter 2

Literature Review

Student attrition has been studied extensively and it dates back to the 1900s by researchers such as [28] and more recently by [2, 18]. The authors explored several factors affecting students. However, there is a growing demand for more advanced ways of analyzing educational data and to incorporate more information [2]. The reasons and potential solutions for student attrition have been investigated in limited quantitative studies [18].

The degree of attrition ranges from institution to institution which is a problem because there would be a lot of loss of resources academically and administratively in addition to negative impact on social level [20, 23].

In this research we adopt the conceptual framework model by Tinto where he relates background, individual attributes, and pre-schooling attributes, to the drop out decisions which is displayed in 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. The input features (a) background or family characteristics, (b) individual attributes, and (c) pre-college attributes impact has been quite explored in previous studies and provide a right prediction for student performance at higher education institutions.

These factors interrelate and influence the student's objective to complete their degree or attitude towards university activities (institutional commitment). In the academic system, creating values and dispositions towards goal commitment translates to improved academic performance and intellectual development. The input observation involved in the Tinto framework deal with biographical and enrollment characteristics [28]. These observations are seen to influence student attrition [2].

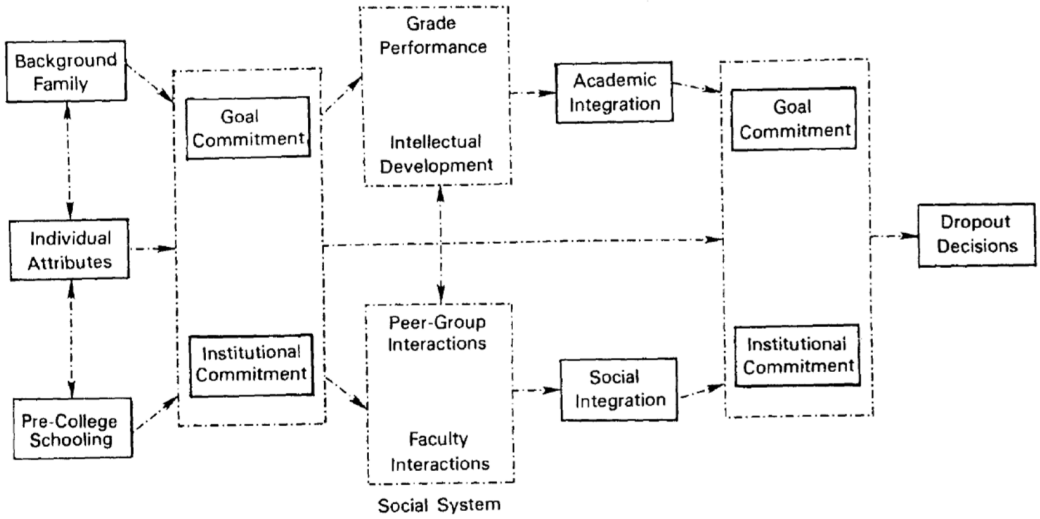


Figure 2.1: Conceptual framework model of Tinto [28] showing the relationship between background characteristics, individual attributes, and pre-schooling attributes to the drop out decisions.

The table in 2.1 shows the use of learners Background, individual and schooling attributes as the best feature-set to predict learner attrition from several authors in previous literature. The first column represents the name of the authors, the second column represents different sets of datasets used, the third column represents features, the fourth column represents the different models explored and the last column represents the various accuracies obtained from the authors.

Table 2.1: A table relating the fundamental factors to review of key authors who used varying feature sets and models to predict learner attrition

Author	Data	Features	Models	Accuracy
Nghe et al. [2007]	Can Tho University (CTU) and the Asian Institute of Technology (AIT).	Demographic variables (gender, age, marital status, area etc), cumulative grade point Average, field of study and English Skill.	DT and BNT	72.95% and 61.54%.
Ajoodha and Jadhav [2019].	Academic Information Systems Unit(AISU) at the University of the Witwatersrand	Biographical characteristics (i.e. gender, spoken home language, home province, and race description) and enrolment observations.	BNT	70%
Sangodiah et al. [2015]	Higher learning institution in Malaysia	Several Background, family and individual characteristics; academic and medical characteristics.	Linear SVM	89.95%
Romero et al. [2008]	Moodle Data.	Final marks obtained in their respective courses.	Applied discretization and rebalance. DT, Rule induction, Fuzzy rule induction and NN.	Less than 70% results for all the models.
Ajoodha [2020]	Under-graduate degree (2008 - 2018), at a South African Higher-Education Institution.	Learners background, individual characteristics, and Grade 12 marks.	NN, DT (C4.5), and probabilistic graphical models.	Accuracies: 85%, 84%, 84%, 83%, 82% respectively.
Kabakchieva and Dorina [2013]	University of National and World Economy data (2007 – 2009).	Pre-university characteristics (gender, birth year, birth place, living place and country, admittance exam and achieved score, university specialty/direction, current semester, total university score, etc.	DT (J48) , Bayesian classifiers, KNN, Rule learners (JRip, OneR).	66.59 %, 60%, 60 %, 63 % and 54-55 %.
Aulck et al. [2019]	University of Washington data (2017).	Students' demographics, complete transcript records and information from applications records.	LR, KNN, RF, SVM's, and gradient boosted trees (XGB).	Accuracies: (83.2%, 83.1%, 83.0%, 82.5% and 78.0% respectively.
Johnson et al. [2015]	Four school districts across the USA.	The datasets contains several attributes; course enrolment and grades, absence rates, tardiness.	RF	79%
Ajoodha and Jadhav [2020].	AISU data (208-2018).	Background features; Individual attributes; Pre-college or schooling	DT, K-Star, Naïve Bayes, SVM, RF and LR.	75%, 64%, 69%, 59%, 74% and 72% respectively.
Osmanbegovic et al. [2012]	Survey data University of Tuzla (2010-2011).	Socio-demographic variables (high school results, entrance exam, and attitudes towards studying).	C4.5, Multilayer Perceptron and Naive Bayes.	73.93%, 71.2% and 76.65% respectively.

2.1 Features

The high number of first generation university students from low-income, less educated families is another problem that leads to the dropout rates [2]. This given background Features (i.e. gender, spoken home language, home province, race description, student performance etc.) have been explored for this problem.

In this section we present the influence of three characteristics (i.e. background attributes, individual variables and schooling attributes) on student attrition which appears in the data. We adopt the conceptual framework model by Tinto where he relates the background of family, individual attributes, and pre-schooling attributes to predict student attrition [28].

Background attributes which includes gender, marital status,age, race description, language, family background, area and qualifications, birth place, living place, spoken home language, home province, and race description have been explored by several authors such as [17], [2], [3], [14], [23], [28] to predict student attrition. While [17] found that the living area/location where international students come from seems to be a barrier, because they come from institutions with diverse grading systems and have backgrounds that faculty and staff are often unfamiliar with. Historical information about each student is an important predictor [1].

Language is another barrier for success in tertiary institutions according to [2]. The relationship between written English skill and academic achievement was examined for the undergraduate programme in an English-medium university [9]. It identified reliable features for timely and cost-effective screening of academically vulnerable students. In the student grade point average (GPA), it is reflected as an academic achievement predictor where measures that were examined were written English skill inclusive of academic reading, academic writing and vocabulary recognition.

In relation to individual attributes, these variables have been explored: medical conditions, academic literacy, quantitative literacy and mathematical literacy, admittance exam characteristics, absence rates, plan code, tardiness and majors [23, 2,

18]. Much research has found that the learners' interests, motivations, study patterns, and family support contribute substantially towards the completion of their degree. Plan code, majors, and chosen science streamline emphasize the value of learners individual attributes [2]. Individual attributes as indicated by [2] indicate the learner's vulnerability more than any background or schooling attribute according to the author.

Pre-college or schooling characteristics have been investigated by [2, 22, 13]. These factors include cumulative grade point average, field of study, final marks, university score and course enrollment. This particular avenue has been explored extensively since most universities base their acceptance criteria solely on an aggregation of the learner's top subjects and data availability [2]. Studies found that features generated from transcript records, aggregates and summaries of students academics are among the factors influencing in the prediction of performance [14].

First-generation students who are vulnerable increases in numbers resulting in entrance to college with multiple risk factors [26]. In line with other studies, most vulnerable students who tended to do worse academically had an unsatisfactory overall educational experience [24].

Not all variables are relevant for a particular context, it is important to conduct detailed studies to identify the context-specific determinants for early interventions to be carried out in a timely manner [2]. It is also necessary to take into account motivational attributes when examining drop-outs in the field of education, including individual characteristics such as school background, academic skills and background characteristics.

2.2 Model and Accuracies

Machine learning paradigms have been used to address attrition in higher education. These include the use of Logistic Regression (LR), K-Nearest Neighbourhood (KNN), Random Forest (RF), Support Vector Machine (SVM), Decision Trees, Gradient Boosted Trees (XGB), multi spectral analysis, Rule induction, Bayesian Networks (BNT), Neural Networks (Multilayer Perceptron), probabilistic graphical

models. It appears that most of the authors found success in using these models with the best accuracy reporting from the literature being the Linear SVM which achieved 89.95% using accuracy as the metric [23].

Several studies have implemented predictive models rather than modelling conventional statistics [22], [20]. There are different types of classification methods and artificial intelligent algorithms used to predict the student results, or ratings [22]. These models have been widely used for predicting drop out. Machine learning algorithms compare how various models of classification can facilitate predictive power.

The authors [17], [22], [3], [2] have applied LR, KNN, RF, SVM predictive modelling whereby, RF produces accuracies of >79% -89% which is the better performing models in comparison to other predictive models.

During the prediction of students performance, various methods and data mining techniques were compared, applying data collected from the surveys at the University of Tuzla, academic years 2010-2011 for first year students [18]. The performance of the learning methods were evaluated based on predictive accuracies of 76.65% for Naive Bayes, 73.93% for Decision trees and 71.2% for Multilayer perceptron.
The precision of the tests was evaluated using GPA cut scores with machine learning methods (logit regression analyzes and Receiver Operating Characteristic (ROC) curves with good accuracies [9].

Student attrition was modelled by [18] using a dataset consisting almost entirely of information collected regularly for record-keeping at a large public university in the USA using one of the largest documented attrition exam datasets (Total Population = 66,060). The results showed the re-enrolment of students for the second year, and subsequent graduation can be predicted reliably based on a single year of data (Area under the Curves = 88% and 81%, respectively). In the field of education data mining, the authors applied the predictive modeling Bayesian approach [2], [17] to their studies with 70% and 61.45% accuracies respectively; compared the performance and convenience of different data mining techniques for student classification using a Moodle mining tool [22].

Chapter 3

Research Methodology

In this chapter we explore the data, methods and research design of the study. The later sections follows in this way: The methods, data (features used to predict the class variables), the brief descriptions of the machine learning classifiers used to perform the predictive and evaluation matrices.

In this section we present the research design used in this report. This study placed its research design according to the context put forward by the Nature and Relevance of the science by [28]. This paper uses a method of Descriptive analysis, since it attempts to construct a model of learner attrition across several possible variables. And this work is quantitative. This report follows the conceptual framework model of Tinto [28] showing the relationship between background characteristics, individual attributes, and pre-schooling attributes to the drop out decisions.
The rationale of this study is that by using background characteristics, individual attributes and Grade 12 marks (outcomes) we can better identify learners with vulnerabilities.

In this report we attempt to answer the following research questions. What are the key potential factors characterizing learners completion of their undergraduate programmes?. What potential features can help us predict learner vulnerability? Which of the adopted classification models are most suitable for classifying learners into risk profiles using these potential factors?

Machine learning predictive models of different archetypes are to be trained. The study uses background, individual and schooling attributes to predict distribution

over risk profiles (response variable) as similar to the Tinto (1975) framework. Classification maps data into predefined groups as classes. The response variable has different Risk Profile namely: Lowest Risk, Medium Risk, High Risk, and Highest Risk. The problem has four classes and 13 variables. Predication can be thought of as classifying an attribute value into one of a set of possible classes.

3.1 Data

Synthetic simulated data learned from Bayesian Networks structure was used in this case. The survey data collection methods included biographical samples, programme characteristics; and assessment ratings from a broad respondent sample.

The synthetic data used in this study consisted of biographical and enrollment observations of learners. It is a Bayesian network generated pro-grammatically. Where, the Bayesian networks gives the underlining ground truth distribution for the variables, therefore it models the conditional dependencies between these variables. The ground truth distribution gives the model that describes what we feel is the relationship between variables. Forward sampling methods were used to sample data instances from the Bayesian Network given in 3.1 using a topological ordering and the structure hypothesized is provided by figure 3.1.

The synthetic dataset used in the study had 41 variables and 50 000 sampled observations. After feature selection, the variables were reduced down to 24. The features were selected by their relevance in terms of our purpose.

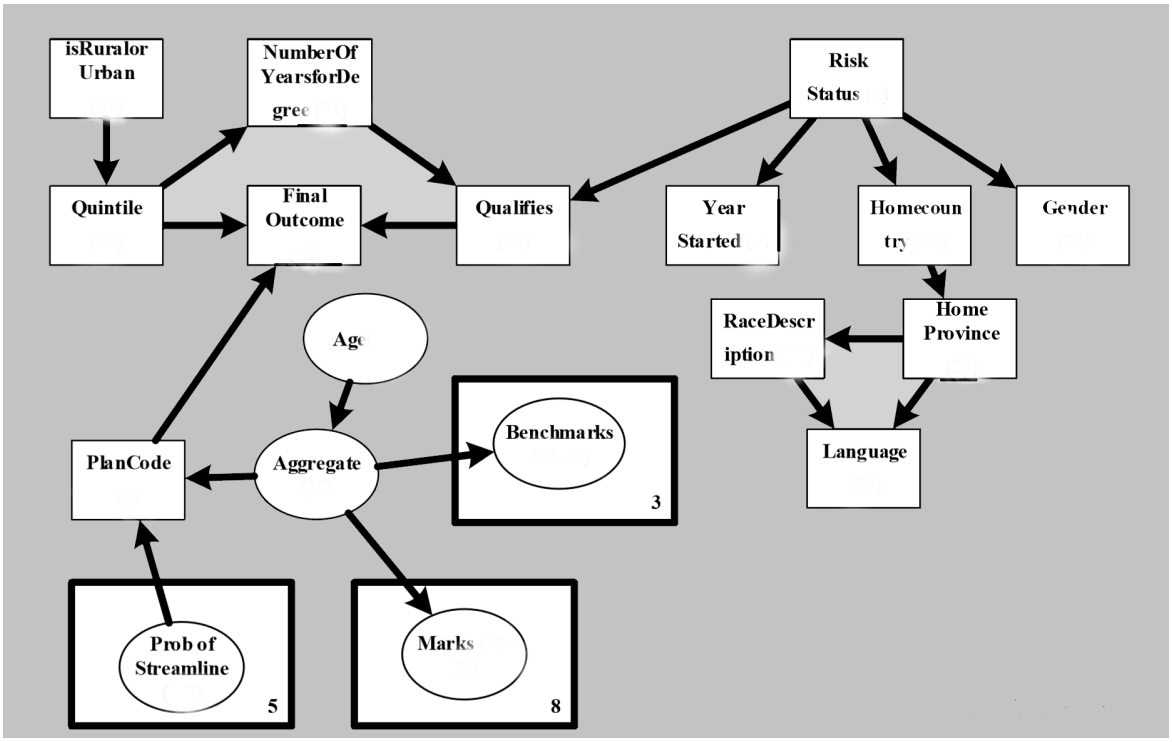


Figure 3.1: This Bayesian Network Structure to predict learner vulnerability

The Bayesian Network Structure to predict learner vulnerability is presented on 3.1. In the diagram, the round variables are continuous and the square variables are discrete. The Gaussian distribution was used to represent the continuous variables. The discrete variables have been computed using tabular conditional probability density (CPD), where you specify the factor levels. The different variables represented as nodes included Risk Status, Plan Code, Prob Of Streamline, Quantile, Aggregate, Progress outcome, Final Outcome, Aggregate, Qualified, Number Of Years for Degree, language, Home Province, Race description, Year started, Home Country, Benchmarks, Marks, Age and Rural or Urban.

Where, aggregate represents: The aggregate of marks 0 to 100 , Qualified whether the student Qualified or Failed, Years in degree 1 to 13,Prob of streamline (maths, physics, earth, biological), The probability of being successful at a particular streamline 0 to 1 , The school location whether urban or rural, School quintile the ranking of schools (quintile) 1 to 5, Marks 0 to 100 , Gender Male or Female, AgeAge of

the students, Race Black, Chinese, Coloured, English and Indian, Languagedifferent languages spoken by students and Home province of the students.

An example of a Conditional independence assumption is given by:
P(The student qualifying)=P(Number of years they took to complete degree) | P(Risk Status)
P(Final outcome) =P(Quantile) | P(Plan code) | P(Whether student qualifies or not?)

3.2 Features

To select the attributes which will enable the researcher to identify learner vulnerability, a list of potential attributes were identified guided by the computed Information Gain (to select the most contributing features) [2].

Students enter a programme [2] with the following Risk status (relating to their possibility of completing their degrees) or outcomes: Lowest Risk (Completes degree in their record specified time); Medium risk (completes their degree in more than the minimum specified time (> 3 years); High (fail to complete their degree in longer than the minimum time (>3 years)) or Highest risk (fails their degree before the minimum time of completion (<3 years) i.e Drop out.

The data used in this study consisted of biographical and enrollment data with Synthetic data. Attributes of the main data are shown in Table 3.1:

Table 3.1: A list of attributes

Attributes		
Biographical information	Academic information	Pre-Schooling attributes
Risk Status, Plan Code, Plan Description, Prob of Streamlines, Progress outcomes, outcomes, Qualified, Grade 12 Marks and Number Of Years for Degree, Aggregate, NBTMA, NBTQL and NBTAL	Race Description, Quintile, Home province, Home country, Rural or Urban, Gender and Language.	English First Lang, English First Additional, Computer Studies, Life Orientation, Mathematics Matric Literacy, Mathematics Matric Major.

3.3 Descriptive Statistics

A description of how different variables varies is analyzed and presented in the next Chapter i.e describing, presenting, summarizing and organizing the data (population). This is described using this statistical distributions, bar graphs,pie charts, ox-and-whisker plots.

Box plots provide a good graphical picture of the concentration of the data. They also explain how far from most of the data the extreme values are. It is a type of graph frequently used to visually illustrate the distribution of numerical data and skewness by showing the quartiles and averages of data in explanatory data analysis.

A Pie Chart is a type of graph that shows a circular graph with details. In each category, the pieces in the graph are proportional to the fraction of the whole, i.e. each slice of the pie is proportional to the size of that category in the whole group.

3.4 Research Models

This section describes the methods to carry out the study. The data explained further in section 3.1.

The following predictive machine learning models are used: Random Forests, Decision trees, Support vector Machines, Bayesian Naive classifier and Multinomial Logistics regression [18]. Classification is probably the most common and familiar technique used in data mining [20].

3.4.1 Random Forests

The random forest classifier is a combination of tree classifiers where each classifier is generated using a random vector that is sampled independently of the input vector and every tree casts a unit vote to identify the input vector in the most common class [19] [13]. Random forests are methods of classification learning that use training data to build multiple decision trees based on the category mode [2]. Each classification tree suits a data bootstrap set, but with the binary partitioning of the tree only a limited number of randomly chosen variables are available at each node [13].

3.4.2 Decision Trees

Decision trees known as supervised classification because of the independent (classes) and dependent attributes given. A Decision Tree or a classification tree is used to learn a classification function. Its a classification method that determines the value of the dependent attribute (variable) based on the values of the independent (input) attributes (variables) is used [5]. The tree intuitively allows the most critical feature to make the decisions from the root down the tree, with respect to the class attribute [2].

3.4.3 Support Vector Machines

Support vector Machines are methods for building a classifier. This method creates a decision boundary between two groups that allows one or more feature vectors to predict labels [11, 13]. The model incorporates the training data into a non-probabilistic linear binary classifier which separates the training data classes through a multi-dimensional hyperplane [2]. The decision boundary known as the hyperplane is orientated in a way that from the closest data points of each class it is as far as possible. Such closest points are called support vectors [11].

Support vector machines are used in most cases for smaller dataset, due to processing time of larger datasets. It has different kernels. The linear and radial basis function kernels are selected for this problem, and the kernels are different in the case where the decision, which is the hyperplane decision boundary between the classes are made. In some (very high-dimensional) feature space, the kernel is a way to compute the dot product of two vectors x and y.

3.4.4 Multinomial Logistic

Multinomial logistic regression is a classification method that extends from the binary logistic regression and rather than consisting of two dependent classes, it consists of more than two dependent or result factor classes. Multinomial logistic regression based on multiple independent variables or the probability of classes of a group, is used to estimate the categorical location of a dependent variable [25]. Classification of multi classes makes the assumption that each sample is allocated to a single label.

3.4.5 Bayesian Network Classifier

A Bayesian network classifier is a Bayesian network used to predict a discrete variable of a class C. It assigns x, an observation of n predictor variables (features) X= (X1; ... ; Xn), to the most probable class: $c* = argmaxcP(c|x) = argmaxcP(x, c)$ [7].

A classifier which assumes strong (Naive) independence assumptions based on Bayes' Theorem is known as Bayesian Network Classifier. For the underlying probability model a more descriptive term would be independent feature model.

This classifier learns the conditional probability of a Ai attribute given the class label C from the training results. Classification is then done by applying the Bayes rule to measure the probability of C given the particular instance of A1; ... ;An, and then to estimate the maximum subsequent probability of a class.

Recent work in supervised learning has shown that a surprisingly simple Bayesian classifier with strong assumptions of independence among features, called naive Bayes, is competitive with state-of-the-art classifiers [7].

3.5 Methods

The programming language for this research is R, R studio and Python 3. They are going to be used for the preprocessing of the data and further analysis.

The dataset will be split into training and test data set [3]. For the purpose of the analysis a 10-fold cross validation is selected. The 10-fold cross validation technique is used to test the algorithm, which is programmed with the same random seed in order to ensure that the same splits are performed on the training and testing data and each algorithm is evaluated exactly the same way. Machine learning algorithms in python scikit-learn will be used to train (75% of the data) and test (25% of test dataset).

To compare the performance of multiple machine learning algorithms consistently, the researcher has to ensure that each algorithm is evaluated in the same way on the same data. Hence a technique called hyper parameter tuning was used. The problem of minimizing a loss function over a graph-structured configuration space is the optimization of hyper-parameters [4].

The Hyper parameter tuning called Grid search will be used. Grid search is a hyper parameter tuning approach that for each combination of algorithm parameters specified, a grid will methodically build and evaluate a model. A grid search creates a grid of hyper-parameters with different combinations and obtain the optimal parameters.

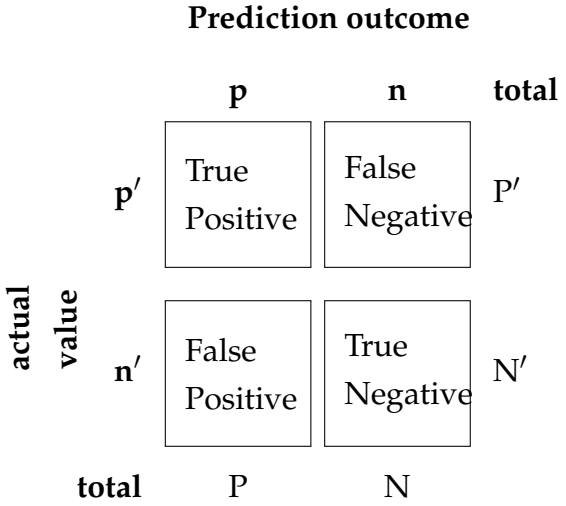
The Grid search will select the best possible combinations of hyper-parameters (cost or gamma values to use) and hence, find the optimal hyper-parameter (the parameter that works best) for the model to train. The Grid Search takes a dictionary describing the parameters that could be attempted to train on a model. The parameter grid is defined as a dictionary, with the keys being the parameters and the values being the settings to be tested.

3.6 Analysis and Evaluation

This section describes the analysis of the data generated in carrying out the research. The steps to follow in the analysis:

3.6.1 Metrics

Data is first described (counts of each response variable). The validation metric for the multi-class classification evaluated are: Accuracy, Recall, and F1-score and precision i.e mean average precision. Confusion matrices were used to gauge the performance of the model [2].



Accuracy is the most intuitive performance measure [12]. Accuracy is a ratio of counts correctly predicted to the total observations.

Precision as mentioned by [12] is the ratio of correctly positive predicted counts to the total positive predicted counts. High precision [12] implies that, they is low false positive rate. That is:

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

F1-score as defined by [12] is the weighted average of Precision and Recall Hence, false positives and negatives rate are taken into account. That is:

$$F_1 = 2 * (\frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}})$$

Recall is the probability that all the given results in the actual classes are positively correctly predicted [12]. That is:

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

3.6.2 Comparisons

The main comparative baselines used for this report are: confusion matrices, model accuracies, recall, precision and f1-scores. Therefore each of this baselines are implemented for these models: Random Forests, Decision trees, K Nearest Neighbors, Different Support vector Machines and multinomial Logistics regression is to be compared [15].

3.6.3 Training/Testing Split

The data was split into two parts both for training and testing i.e. 75% for training and 25% for testing. We used splitting by target variable (risk status) into equal class proportions.

We used k-fold cross-validation for evaluating our model results, with k =10 folds where k is the number of groups to divide the data. The data is then partitioned randomly into similarly spaced k subsets. In addition, each subset is used to test the model installed on the rest of k-1 subset.

3.6.4 Feature Selection

We will use Information Gain Ranking (IGR) algorithm to perform feature analysis. The IGR algorithm calculates the entropy (information gain) in respect of the class variable. The entropy, $0 \leq e \leq 1$, where 0 indicates no information gain and 1 indicates maximum information gain. In the next section we will describe the machine learning results for the classifiers used in this report. 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).

3.7 Limitations

The study, after receiving data from the Academic Information System Unit, encountered problems in terms of using the received data due to procedures pertaining privacy of the students. Hence, synthetic data was used to conduct the study. In addition, the size of the theoretical data used for the research had possible limitation of generalizing the study. While, due to time constraints, machine learning methods such as Multilayer Perceptron were not tested.

Chapter 4

Results and Discussion

This chapter presents the analysis in relation to the purpose and objectives of the study. Furthermore, the chapter presents the interpretation and discussion of the research findings.

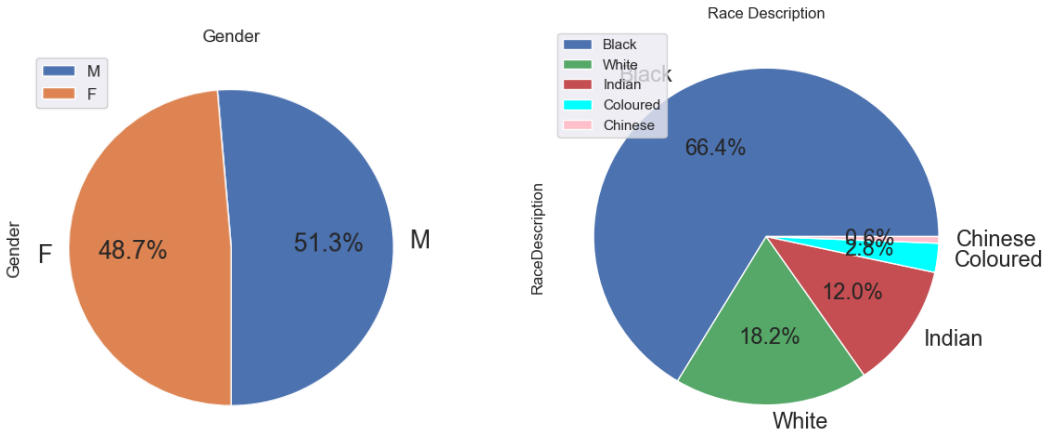
4.1 Analysis of data

The synthetic data of the academic years of 2010-2018 were analysed on Python 3, R, Matlab 2018b and weka 3.8 and results are presented with respect to the variables. The machine learning technique mentioned in Chapter 3 are used to display the results in terms of the description of the different machine learning models.

Students enters a programme, the following Risk status (relating to their possibility of completing their degrees) outcomes are possible: No Risk (Completes degree in their record specified time); Medium risk (completes their degree in more than the minimum specified time (> 3 years); High (fail to complete their degree in longer than the minimum time (>3 years)). or Highest risk (fails their degree before the minimum time of completion (<3 years) i.e Drop out.

4.1.1 Descriptive statistics

We explore how the independent variables relate to each other and to the target variable (risk status).



(a) Gender **(b)** Race Description

Figure 4.1: The pie chart of the gender and race description of the learner.

In figure 4.1, we describe the distribution of Gender. The proportions of the gender s Male is (51%) and Females (49%). There is not much variation in Gender. We also describe the distribution of race. The proportions in race shows that the highest proportion is black (66%); followed by white (18%); then Indian (12%); then coloured (3%); and the least is Chinese roughly(1%).

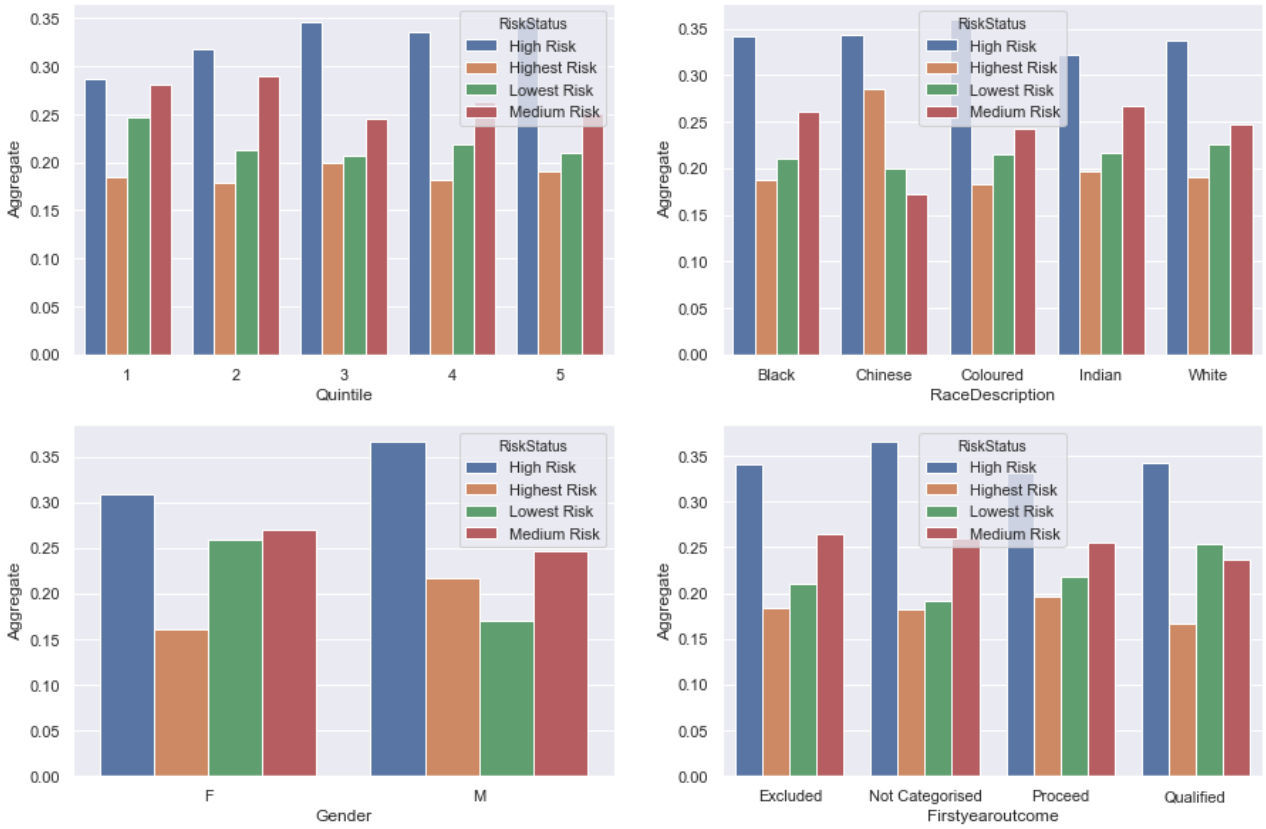


Figure 4.2: The distribution showing the relationship between Risk Status and different variables: Quintile of the learners, Race Description, Gender and first year outcomes .

Figure 4.2 shows the distribution of the different relationships between the Risk status and the variables Quintile of the learners, Race Description, Gender and First Year Outcome of the learners. The distributions with respect to the risk status shows a variation for each class whereby High risk class is contributing the most and the classes Highest risk and Lowest risk have minimum contribution.

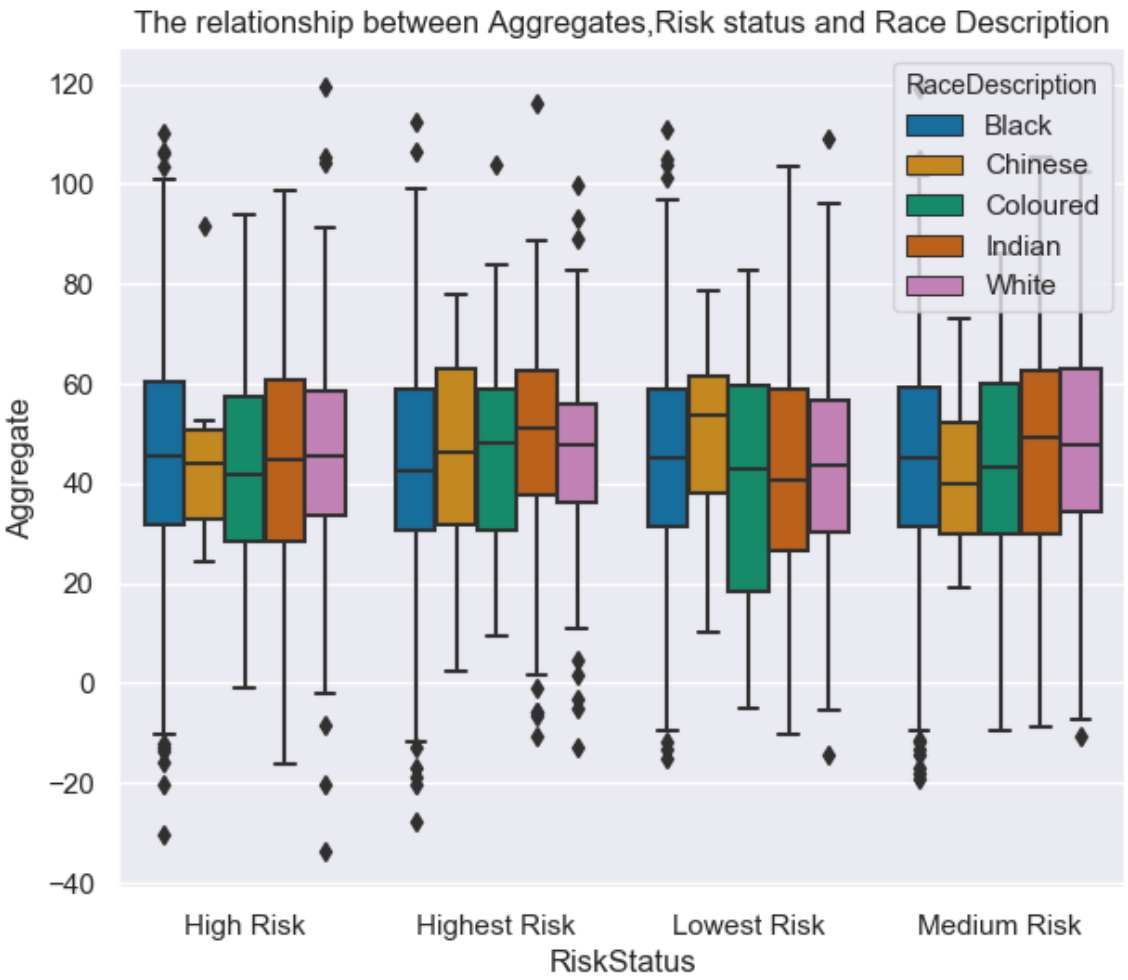


Figure 4.3: The relationship between Aggregate, Race Description and the Risk Status.

The study has 5 race descriptions with their respective proportions: Black, Chinese, Coloured, Indian and White. For the risk profiles high risk, Black learners perform better (i.e have higher aggregates); while the risk profiles highest risk has Chinese and Indian learners who perform better; While the racial group White has a higher aggregates for the Medium risk and for the Lowest risk profile, Chinese and Blacks perform better. The learners with the Lowest risk profile has overall higher aggregates for all race descriptions.

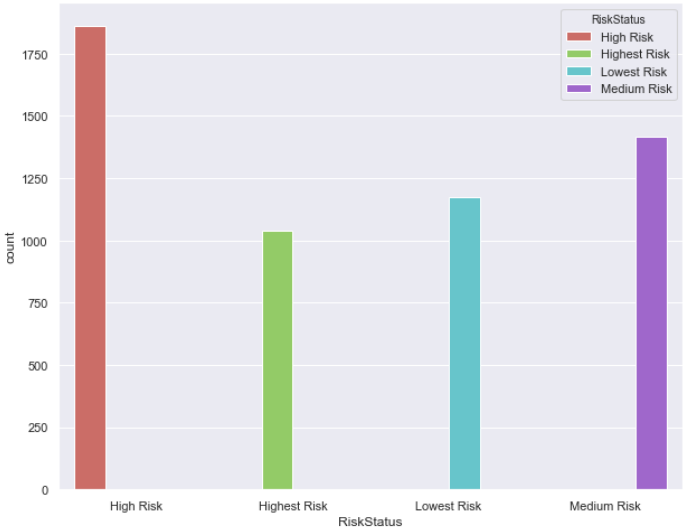


Figure 4.4: Total count of each response variable (Risk Status). The response variable represents: Lowest Risk, Medium Risk, High Risk and Highest Risk.

The target variable risk profile has four risk profile: lowest risk profile with the student expectation of completing the degree in record time (three years); medium risk profile with the student expected to complete the degree in more than the minimum time; the high risk profile with the student expected to compete the degree after after a long time; and the highest risk profile where the student is expected to not 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). The purpose of the study is to classify or deduce the students into four risk profiles. 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: largest proportion is high risk profile with (34%); followed by medium risk profile (26%); then the lowest risk profile (21%); and the least proportion is the highest risk profile (19%).

4.1.2 Feature Information Gain

This section explores the contribution of each of the features to classify the class variable given by risk status using IGR. Table 4.1 illustrates a ranking of the contribution of each feature to classify the Risk Profile using IGR.The first column (Rank) shows the ranking of features from the most significantly contributing features with a high IGR to the least contributing feature with the lowest IGR. The second column represents the Information Gain (entropy). The last column is the feature name associated with the ranking. The IG is the value between $0 \leq e \leq 1$ whereby 0 shows that there is no information gain/contribution, and 1 highest IG. The features are color coded relating them to the conceptual framework by Tinto where red indicates background characteristics, blue indicating individual attributes, and black indicating pre-college or schooling data.

The eleven most contributing features are the following: English First Lang and English First Additional; Plan Description, Year Started, Number Of Years for Degree, Qualified, Progress outcome and Plan Code; and Home province, Quintile, Race Description. The top features suggest that biographical characteristics are the most dominant in deducing the student risk status followed by individual attributes as per the Tinto framework [28]. The variable pre-college attributes show no or minimal effect on student risk profiles. Understanding the role that these factors play can help us uncover cues to expedite learner retention and progression and thus degree completion.

Table 4.1: A ranking of the information gain (entropy, denoted e) for a set of features to predict learners Risk Profile. The top 11 features are highlighted in light blue.

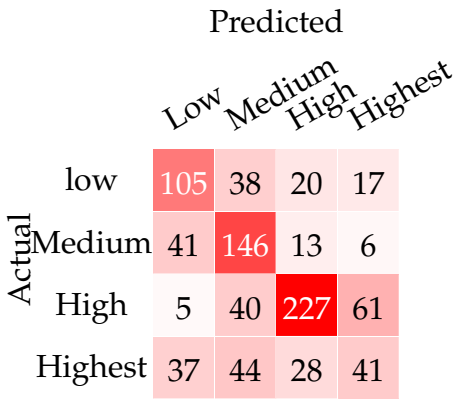
Rank	e	**Feature Name**
1	1.36	English First Lang
2	1.32	English First Additional
3	1.23	Plan Description
4	1.18	Number Of Years for Degree
5	1.10	Race Description
6	1.10	Year Started
7	1.01	Progress outcome
8	0.88	Quintile
9	0.84	Qualified
10	0.73	Plan Code
11	0.24	Home province
11	0.07	Aggregate
12	0.03	NBTAL
13	0.03	Computers
14	0.02	Life Orientation
15	0.02	Mathematics Matric Lit
16	0.00	First Year Outcome
18	0.00	Mathematics Matric Major
19	0.00	Rural or Urban
20	0.00	NBTMA
21	0.00	Gender
22	0.00	NBTQL
23	0.00	Language
24	0.00	Home Country

4.1.3 Results of models: Confusion Matrix and Accuracies

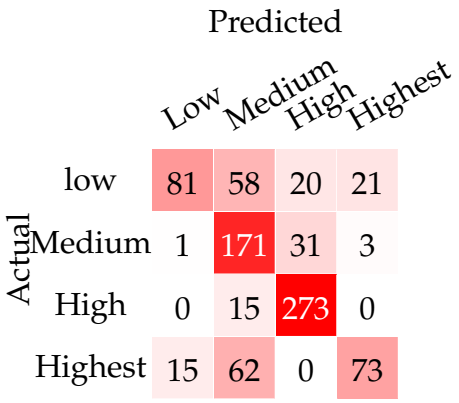
In this section, the results of the prediction models are presented. The following five predictive procedures were employed in this report: Random Forests, Decision tree, Support Vector Machines, Bayesian Classifier and Multinomial Logistics Regression. Figures below indicates the result of each of these classifiers to predict the class variable. The following tables provide the description of the confusion matrices for all the fitted models and their respective predictive performances.

Confusion Matrices:

This section displays results of models:



(a) A confusion matrix describing the performance of the **Multinomial Logistic Regression** predictive model.

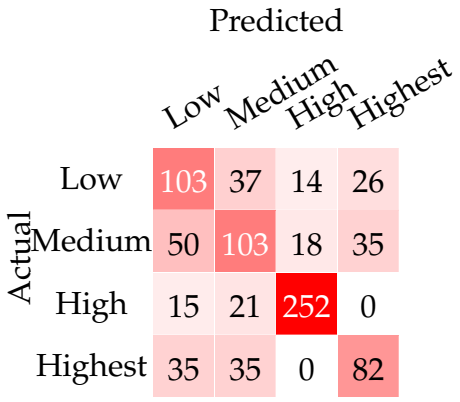


(b) A confusion matrix describing the performance of the **Random Forest** predictive model.

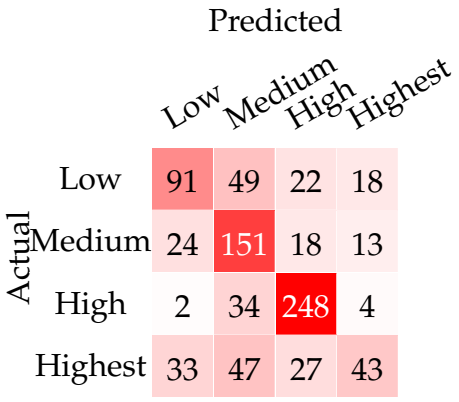
Figure 4.5a shows the confusion matrix describing the performance of the Multinomial Logistic Regression predictive model. The model struggles to detect the lowest risk class and the highest risk class, this may be due to their class proportions of the low risk being slightly lowest: 21% ad 26% respectively. The lowest risk and highest risk classes are often falsely (misclassified) as Medium risk class. The Multinomial Logistic regression predictive model achieves accuracy of 63% with Best Hyper Parameters using the 10-fold cross validation.

Figure 4.5b shows the confusion matrix describing the performance of the Random Forest predictive model. The Random Forest classifier best predicts or classifies

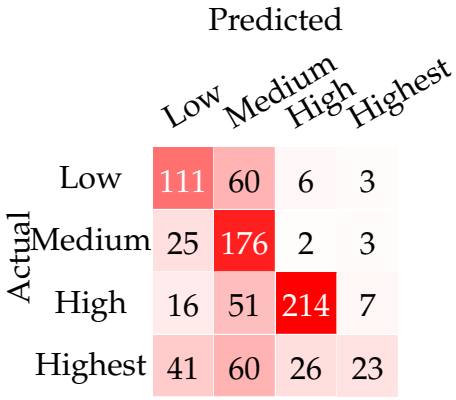
medium risk class and high risk. The model often classifies Highest risk class as Medium Risk. Random forest predictive model predictive model achieves accuracy of 73% with Best Hyper Parameters using the 10-fold cross validation.



(a) A confusion matrix describing the performance of the **Decision Tree** predictive model.



(b) A confusion matrix describing the performance of the **SVM** predictive model.



(c) A confusion matrix describing the performance of the **Bayesian Network Classifier** predictive model.

Figure 4.6a shows the confusion matrix describing the performance of the Decision Tree predictive model. This model often classifies Low and High Risk as Medium Risk. The Decision Tree predictive model achieves accuracy of 66% using the 10-fold cross validation. This model is the worst performing model.

Figure 4.6b shows the confusion matrix describing the performance of the SVM predictive model. This model often classifies Low and High Risk as Medium Risk also. The result of table indicates that the model correctly mis-classifies Low risk

and Highest risk. The Linear Support Vector Machines predictive model achieves accuracy of 65% using the 10-fold cross validation.

Figure 4.6c A confusion matrix describing the performance of the Bayesian Network Classifier predictive model. The model often classifies all the other risk types as Medium Class. The predictive model achieves accuracy of 64% using 10-fold cross validation. Medium Risk is the class that is correctly classified the most by this model. This is the second lowest performing model using the 10-fold cross validation.

The confusion matrices illustrates the confusion matrix for the predictive models. The Medium Risk and High Risk classes are the most correctly classified for each model while Lowest risk and highest risk are incorrectly classified for this report.

Model Accuracies, Recall and Precision

This section displays results of the predictive accuracies of the six trained models. The accuracy was evaluated using 10-fold cross validation method. Table 4.2 describes the results of the fitted models: Random forests, Multinomial Logistic Regression, Support Vector Machines, Bayesian Network and Decision Tree.

Table 4.2: The predictive accuracy of the six trained models.

Model	Accuracies
Random Forest	73%
Multinomial Logistic Regression	63%
SVM	65%
Bayesian Classifier	64%
Decision Trees	65%

Comparing model by model it is clear that the predictive accuracy of the Random forest model is higher than the accuracy of all the other models. Random forest is the best model because it has the highest predictive accuracy using the testing dataset. It correctly classify the instances from the predictions made and it achieves an accuracy of 73%. The Decision tree and the Support Vector Machine achieves the second highest predictive accuracy of 65% following the random forest model. The Bayesian Network Classifier has the third highest predictive accuracy of 64% after random forest and both the Decision Tree and the Support Vector Machines. The Multinomial Logistic Regression has the forth high predictive accuracy of 63% after random forest, Decision Trees, Support Vector Machine and the Bayesian Network classifier.

Recall

Table 4.3: The Recall of the predictive Models

Model	Lowest	Medium	High	Highest
Random Forest	0.45	0.83	0.95	0.49
Multinomial Logistic Regression	0.58	0.71	0.79	0.27
SVM	0.51	0.73	0.86	0.29
Bayesian Classifier	0.62	0.85	0.80	0.25
Decision Trees	0.57	0.50	0.88	0.55

Table 4.3 illustrates the recall metric for the six trained models and for each risk profile class. This will help us describe which models correctly classify risk profiles, and which risk profile classes have higher proportion of correctly classified labels. The recall is the measure of the proportion of actual positives that are correctly classified. The recall was computed from the Classification report given by each model.

The model that has the best overall recall is the Random Forest (have higher proportion of classes correctly classified as their true label), followed by Decision Tree and the Support Vector Machine, Bayesian Network Classifier then the Multinomial Logistic Regression (have lower proportion of classes correctly classified as their true label.) with the least overall recall. Describing the recall rate by classes shows that the high risk profile class has the highest recall most of the observations labeled as high risk class are actually high risk profiles). The lowest risk profile class has the highest mis-classification (implying that the observations classified as lowest risk profiles are actually not lowest risk label) rate across all the models;and the high risk class has the greatest classification rate.

Precision

Table 4.4: The Precision of the predictive Models

Model	Lowest	Medium	High	Highest
Random Forest	0.84	0.56	0.84	0.75
Multinomial Logistic Regression	0.56	0.54	0.79	0.51
SVM	0.61	0.54	0.79	0.55
Bayesian Classifier	0.58	0.51	0.86	0.64
Decision Trees	0.51	0.53	0.89	0.57

Table 4.4 illustrates the precision metric for the six trained models, and for each risk profile class. The precision refers to the percentage of the results which are relevant (prediction of the true class). Table 4.4 illustrates the precision metric for the six trained models, and for each risk profile class. The higher the precision value, the better the model is at predicting relevant risk profiles often.

The model that has the best overall precision is the Random Forest (have higher proportion of classes correctly classified as their true label), followed by Multinomial Logistic Regression and the Support Vector Machine, Bayesian Network classifier then Decision Tree (have lower proportion of classes correctly classified as their true label) with the least overall precision.

F1-score

Table 4.5: The F1-score of the predictive Models

Model	Lowest	Medium	High	Highest
Random Forest	0.58	0.67	0.89	0.59
Multinomial Logistic Regression	0.57	0.62	0.79	0.36
SVM	0.55	0.62	0.82	0.38
Bayesian Classifier	0.60	0.64	0.80	0.25
Decision Trees	0.54	0.51	0.88	0.56

Table 4.5 illustrates the F1-scores for the six trained models, and for each risk profile class. The F1-score is a harmonic mean of precision and recall, that maximizes the F1-score, and there by maximising both precision and recall. The model that has the best overall F1-score is the Random Forest, followed by Decision Tree, the Support Vector Machine, Bayesian Network classifier then the Multinomial Logistic Regression with the least overall F1-score.

4.1.4 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. The purpose of this research was to explore the relationship between background, individual and pre-college characteristics on attrition as per the Tinto model of learner attrition [28]. These characteristics are then used as input attributes to predict the student attrition by classifying the students into four risk profiles.

The pre-college characteristics show minimal effect on deducing student risk profiles. Similar results were achieved by several researchers [2], [23], [17], [9],[6] and

[3] which indicates that biographical and individual characteristics play a major role in deducing students into the correct risk profiles.
The model that performs the greatest with the given features is the Random Forest model with a 73% accuracy supported by the recall, precision and f1 scores over the four risk profiles over all the other models. The poor model is the Multinomial Logistic Regression with accuracy of 63%. Other Models predicted the Risk profiles better.

The Random Forest model with 73% success may be due to the quality of the training data. While the similarity of the performance given by Decision tree and Support vector Machine may be due to the classes and the selected variables.

A shrinking amount of resources and attention is being given to treat learner attrition due to the increased intake of learners at different institutions [2]. It is therefore important to accurately forecast risk profiles for learners through an Early Warning System. In particular, predicting when learners will encounter flaws in their chosen curriculum helps the university to intervene early in order to prevent the learner from dropping out.

4.1.5 Summary

In this chapter we have discussed the results obtained from applying our methods discussed in chapter 3. The results describes output from our trained models. We found important or most significant features in classifying risk status. The top 11 important features provide evidence that biographical characteristics are the most dominant in deducing the student risk status. The Random Forest model outperforms the other models with accuracy of 73%.

Chapter 5

Conclusions and Future Work

This final chapter presents on the main conclusion (brief overview) of the study in relation to the research problem. Future work of the study is also presented.

5.1 Conclusion

This study presents a discussion about student attrition using background, individual, and schooling attributes as per Tinto framework [28]. The significance of this paper was to define potential factors that can be used to predict failure of learners in order to resolve bad results. Information gain shows that deducing the student into the correct risk profile is dominantly affected by background characteristics, followed by individual attributes while the pre-college (schooling) characteristics show minimal or no effect on student attrition. Different research models were explored. In other contexts, the analysis in this paper should be duplicated to evaluate the possible factors that affect learner vulnerability, as the large causal factors associated with learner vulnerability can not be easily established [2].

This simulation is inline with current research on real data as showed by researchers such as [21, 2].

A decreasing amount of funding and attention are being provided to treat learner insecurity due to the expanded intake of learners at many South African universities. It is important to automatically forecast the vulnerability of the learner through

a Early Warning System. In fact, forecasting when learners will encounter vulnerabilities in their chosen program helps the university to interfere early, thereby preventing the learner from dropping out.

The results indicates that the machine learning models employed were able to predict learner vulnerability with the given attributes. The random forest model performed better compared to the other fitted models across all the risk profile or classes; with an accuracy of 73%, followed by Decision Trees with 65%, Support vector Machines with 65% accuracy and the Bayesian classifier with 64%. The least performing model is a Multinomial Logistic regression with 63% accuracy.

Random forest was successful in this report due to the nature of how random forest is in line with how the method handles variables fast, making it suitable for different tasks. This approach generates as many trees on the subset of the data and averages the output of all the trees [13]. In this way it eliminates overfitting problem in decision trees and also reduces the variance and thus improves the accuracy. The results of Random Forest is also supported by research done by several authors [13], [3]. The model Multinomial Logistic regression did not perform the best when using the sets of features from the synthetic data.

Students vulnerability can be predicted using random forest model. Hence, as a comparative study, the mentioned techniques can be used in accordance to this data. These framework to address vulnerability can be used to re-mediate vulnerable students status thereby increasing pass-rates; lowering drop-outs.

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 enrollment system.

The major contribution of this paper: Comparison of different machine learning models by contextualizing the students background, individual and enrollment data to find the better method to analyze vulnerability.

5.2 Future work

In line with the research, this work can be extended in multiple ways:

The study of student attrition is important in educational research. This study used synthetic data hence working with actual real student enrollment data in the future is imperative. This will help us to verify if our theoretical model is applicable in real-world data.

Future work is also inclusive of a) research that focuses on all faculties using University data for vulnerable students; b) add more attributes (variables) to the models such as motivation, determination, and commitment [2]; c) extends to more machine learning techniques to test for learner vulnerability (to obtain better accuracy); d) build a framework where we deploy our algorithm, so that when a student inputs background, individual, and schooling attributes, the algorithm will generate a forecast that shows the risk profile of the applicant.

Bibliography

[1] Everaldo Aguiar et al. “Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time”. In: *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. 2015, pp. 93–102.

[2] Ajoodha Ritesh et al. “Forecasting learner attrition for student success at a south African University.” In: *Association fo Computing Machinery* 1 (2020), pp. 1–10.

[3] Lovenoor Aulck et al. “Mining University Registrar Records to Predict First-Year Undergraduate Attrition.” In: *International Educational Data Mining Society* (2019).

[4] James S Bergstra et al. “Algorithms for hyper-parameter optimization”. In: *Advances in neural information processing systems*. 2011, pp. 2546–2554.

[5] Neeraj Bhargava et al. “Decision tree analysis on j48 algorithm for data mining”. In: *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering* 3.6 (2013).

[6] 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).

[7] Nir Friedman, Dan Geiger, and Moises Goldszmidt. “Bayesian network classifiers”. In: *Machine learning* 29.2-3 (1997), pp. 131–163.

[8] Leonid Grebennikov and Ivan Skaines. “University of Western Sydney students at risk: Profile and opportunities for change.” In: *Journal of Institutional Research* 14.1 (2008), pp. 58–70.

[9] Michael Harrington and Thomas Roche. “Identifying academically at-risk students in an English-as-a-Lingua-Franca university setting”. In: *Journal of English for Academic Purposes* 15 (2014), pp. 37–47.

[10] Joann Horton. “Identifying at-risk factors that affect college student success”. In: *International Journal of Process Education* 7.1 (2015), pp. 83–101.

[11] Shujun Huang et al. “Applications of support vector machine (SVM) learning in cancer genomics”. In: *Cancer Genomics-Proteomics* 15.1 (2018), pp. 41–51.

[12] Renuka J. *âHowtoevaluatetheperformanceofamodelinAzureMLand understanding âConfusion Metricsââ*. 2016. URL: http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performancemeasures (visited on 09/09/2019).

[13] 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.

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

[15] Himabindu Lakkaraju et al. “A machine learning framework to identify students at risk of adverse academic outcomes”. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM. 2015, pp. 1909–1918.

[16] Andy Liaw, Matthew Wiener, et al. “Classification and regression by randomForest”. In: *R news* 2.3 (2002), pp. 18–22.

[17] 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.

[18] 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.

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

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

[21] David C Rheinheimer et al. "Tutoring: A Support Strategy for At-Risk Students." In: *Learning Assistance Review* 15.1 (2010), pp. 23–34.

[22] Cristóbal Romero et al. "Data mining algorithms to classify students". In: *Educational data mining 2008*. 2008.

[23] Anbuselvan Sangodiah et al. "Minimizing student attrition in higher learning institutions in Malaysia using support vector machine." In: *Journal of Theoretical & Applied Information Technology* 71.3 (2015).

[24] Annalina Sarra, Lara Fontanella, and Simone Di Zio. "Identifying students at risk of academic failure within the educational data mining framework". In: *Social Indicators Research* (2018), pp. 1–20.

[25] Jon Starkweather and Amanda Kay Moske. "Multinomial logistic regression". In: *http://www. unt. edu/rss/class/Jon/Benchmarks /MLR_JDS_Aug2011.pdf* 29 (2011), pp. 2825–2830.

[26] Shannon M Suldo et al. "Identifying high school freshmen with signs of emotional or academic risk: Screening methods appropriate for students in accelerated courses". In: *School Mental Health* (2019), pp. 1–18.

[27] Hilary Tait and Noel Entwistle. "Identifying students at risk through ineffective study strategies". In: *Higher education* 31.1 (1996), pp. 97–116.

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