Predictive working tool for early identification of 'at risk' students

Zlatko J. Kovačić, Associate Professor
Phone: + 64 4 913 5777
Email: Zlatko.Kovacic@openpolytechnic.ac.nz

John Steven Green, Senior Lecturer
Phone: + 64 4 913 5724
Email: John.Green@openpolytechnic.ac.nz

School of Information and Social Sciences, Open Polytechnic
Private Bag 31914, Lower Hutt, New Zealand

This version: 20 July 2010

An Ako Aotearoa publication.
Support for this work was provided by Ako Aotearoa through its Regional Hub Project Funding scheme

Executive summary

This paper explores the variables that may influence persistence or dropout of students at the Open Polytechnic. These include socio-demographic variables such as age, gender, ethnicity, education, work status, and disability as well as variables related to the study environment such as course faculty (School of Business, School of Information and Social Sciences and Workplace Learning and Development), programme (Bachelor of Business, Bachelor of Applied Science and Bachelor of Arts), level (Level 5, 6 and 7), block (Trimester 1, Trimester 2 and Trimester 3) and offer type (Distance, Blended and Online).

We sought to determine the extent to which data captured by the enrolment form could help us to identify future successful and unsuccessful students before the course began. This would enable us to provide guidance to students on their course choices and to be able to focus additional support on those students statistically more likely to fail. All too often students enrol on courses at a level too high for their current skills; find themselves at risk of failing.

Data from 2006 to 2009, covering over 19,400 enrolled students stored in the Open Polytechnic student management system was used to perform a quantitative analysis of study outcome. Using various data mining techniques the most important factors for student success were identified and typical profiles of successful and unsuccessful students were constructed. For the Open Polytechnic, the student most likely to be successful is European with University Entrance or an overseas qualification and female and will pass with a probability of 0.921. The student with the greatest number of indicators of failing are either Māori or Pacific Island studying a level 5 course in the Bachelor of Applied Science. They will fail with a probability of 0.751.

The empirical results show that the most important factors separating successful from unsuccessful students in order of importance, were: ethnicity, course level, secondary school qualification (highest level of achievement held from a secondary school), programme and age.

- *Ethnicity* is not something a student can change or an institution can influence, but advice on the most appropriate study options for that student, i.e. distance, online or contact study may be provided. Would some students be better served by studying in a contact institution, or if in a contact institution by distance?
- The factor *course level* is deceiving, as it might suggest that students studying a lower level course are more likely to succeed. In fact the reverse is true. Students on higher level courses who have already proven themselves in lower level courses are more likely to succeed, making this a relatively predictable result in much the same space as the third factor, secondary school qualification.
- *Previous academic success* is a strong indicator of future academic success and has been used in the UK by University Matriculation Boards for decades.

- Advising a student that a different degree *programme* might increase their chances of academic success appears to be fraught with difficulties but in a larger institution with more choices of programme it may be quite appropriate.
- Like ethnicity, *age* is not something a student or institution can modify but it should come as no surprise that younger less mature people have less motivation than older more mature people who generally have a higher motivation to succeed.

The implications of these results for academic and administrative staff are several. The implications of identifying a student as potentially unsuccessful must be considered. If the student is told how they have been categorised what effect might this have on their self-esteem and subsequent motivation? In tough economic times should the organisation refuse to enrol students statistically unlikely to pass the course? Or should they allocate further resources to support those students with no guarantee that this support will be effective?

Classification using discriminant functions was the most accurate overall but required the consideration of more factors and was less accurate in identifying 'at risk' students. The CART classification tree was the most accurate. Regardless of the method used, our results suggest that using enrolment data alone is only moderately successful in separating successful from unsuccessful students.

It is essential to recognise that while this model will effectively separate successful and unsuccessful students with a good level of accuracy the results are specific to the population analysed. The results would need to be regularly updated with each passing trimester and for a different student population. This would allow each unique student body to be modelled and a checklist used to identify potentially unsuccessful students prior to enrolment.

This study is limited in the three main ways that future research can perhaps address. Firstly, our research is based on enrolment data only. Leaving out other important factors (academic achievement, number of courses completed, motivation, financial aids, etc.) that may affect study outcome could distort results obtained with models used. For example, including the assignment mark after the submission of the first course assignment or even better a pre-entry test would probably improve the predictive accuracy of the models. To improve the model, more attributes could be included to obtain prediction models with lower misclassification errors. However, the model in this case would not be a tool for pre-enrolment, i.e. early identification of 'at risk' students.

Secondly, the time line should be included in the analysis. We would need to follow those students who failed the course and also transferees and withdrawal students. Some of them may re-enrol in one of the next semesters and might successfully complete the course at the second or third attempt. Tracking *Fail* and *Lost* students in subsequent semesters and tracking their study outcomes would help make modelling their behaviour more accurate.

Thirdly, from a methodological point of view an alternative to logistic regression and discriminant analysis should be considered. The prime candidate to be used with this data set is neural networks. We may also consider other classification tree models such as exhaustive CHAID, QUEST, random forest, and ensembles of models.

Table of Contents

Tables

Figures

1. Introduction

Increasing student retention or persistence is a long term goal in all academic institutions. The consequences of student attrition are significant for students, academic and administrative staff. The importance of this issue for students is obvious: school leavers are more likely to earn less than those who graduated. Since one of the likely criteria for government funding in the tertiary education environment in New Zealand is the retention rate, both academic and administrative staff are under pressure to come up with strategies that could increase retention rates on their courses and programmes.

The lowest student retention rates at all institutions of higher education are first-year students, who are at greatest risk of dropping out in the first term or semester of study or not completing their programme/degree. Therefore most retention studies address the retention of first-year students (e.g. Horstmanshof & Zimitat, 2007; Ishitani, 2003, 2006; Noble, Flynn, Lee & Hilton, 2007; Pratt & Skaggs, 1989; Strayhorn, 2009). Consequently, the early identification of vulnerable students who are prone to drop their courses is crucial for the success of any retention strategy. This would allow educational institutions to undertake timely and pro-active measures. Once identified, these ‘at risk’ students can be then targeted with academic and administrative support to increase their chance of staying on the course.

A number of theoretical models have been developed to explain what keeps students on a course. Based on an extensive literature review of dropout in an e-learning environment Jun (2005) identified variables that may impact attrition and have been included in theoretical models of dropout. He classified them into five constructs, i.e. factors: individual background, motivation, academic integration, social integration and technological support.

Background characteristics such as academic and socio-demographic variables (age, sex, ethnic origin, marital status, and financial aid) have been identified in retention literature as potential predictor variables of dropout. Pascarella, Duby, and Iverson (1983) stated that the students’ characteristics are a factor of equal if not greater importance when deciding to stay or discontinue the study, than the actual experience once enrolled. In Bean and Metzner's (1985) conceptual model of non-traditional student attrition a set of background characteristics is causally linked to the effect that academic and environmental variables have on the outcome of persistence or dropout. As Tharp (1998) stated after an extensive literature review, that the background characteristics taken alone as predictors of dropout have not performed well in the case of traditional students (regular, full-time students). However, the background information was significant in the case of non-traditional students (distance/open education) where social integration and institutional commitment are not central to the student experience.

Studies by Jun (2005) and Herrera (2006) provide a comprehensive overview of the theoretical models describing student persistence and dropout in both contact and distance education institutions. Grote (2000) also provided an overview of earlier literature on student retention and support in open and distance learning concluding pessimistically that modelling retention data “are likely to remain unsuccessful”. Traditionally, from the methodological point of view, statistical models such as logistic regression (e.g. Glynn, Sauer & Miller, 2003; Woodman, 2001) and discriminant analysis (e.g. Dirkx & Jha, 1994; Dupin-Bryant, 2004) were used most frequently in retention studies to identify factors and their contributions to student dropout. There are also other, less frequently used models such as survival or failure-time analysis (Murtaugh, Burns & Schuster, 1999), and the Markov student-flow model (Herrera 2006) that were used to monitor students' progression from the first to the final year of their study.

However, in the last 15 years educational data mining emerged as a new application area for data mining, becoming well established with its own journal (Journal of Educational Data Mining). Romero & Ventura (2007) provided a survey of educational data mining from 1995-2005 and Baker & Yacef (2009) extended their survey covering the latest developments up to 2009. There are an increasing number of data mining applications in education, from enrolment management, graduation, academic performance, gifted education, web-based education, retention and other areas (Nandeshwar & Chandhari, 2009). In this section we will only review research where the main focus is on study outcome, i.e. successful or unsuccessful course completion.

Based on his open learning model Kember (1995) stated that entry, i.e. background characteristics are not good predictors of final outcomes because they are just a starting point and there are other factors that may contribute to the difficulties a student will have to deal with during his or her study.

Bathurst (2004) reported results of an analysis of Diploma of Health and Human Behaviour completions in 2002 at the Open Polytechnic. By using simple descriptive statistics of demographic data he identified factors that contribute to completion rates. The following categories of students are identified as 'at risk' students: male, Māori and Pacific Islanders, and those with minimal or no secondary school qualifications.

Woodman (2001) found that for courses in the mathematics and computing faculty at the Open University in UK, by using the binary logistic regression, the most significant factors contributing to whether students passed, failed or dropped out, were the marks for the first assignment, the number of maths courses passed in the previous two years, the course level, the points the course was worth and the occupation group of the student. This was the most parsimonious model, but in the other model which includes all 25 potential predictors, other variables such as ethnicity (ranked as 7th according to its relative importance), education (8th), age group (9th), course level (11th), disability (18th) and gender (22nd) were also significant. However, one of the problems with logistic regression when used in large samples is that any small difference could be identified as statistically significant, which may lead to the conclusion that the related factor is significant when in the true, unknown regression model we are estimating, this is not the case.

Using the same methodological approach with data available at new student registration in the UK Open University, Simpson (2006) found that the most important factor is the course level, followed by the credit rating of a course, previous education, course programme, socio-economic status, gender and age.

Kotsiantis, Pierrakeas & Pintelas (2004) used key demographic variables and assignment marks in supervised machine learning algorithms (decision trees, artificial neural networks, naïve Bayes classifier, instance-based learning, logistic regression and support vector machines) to predict a student's performance at the Hellenic Open University of Greece. When only the demographic variables were used the prediction accuracy varied from 58.84% (when using a neural network) to 64.47% (when using support vector machines). However, when other variables beside demographic were included, the naïve Bayes classifier was found to be the most accurate algorithm for predicting students' performance.

Vandamme, Meskens & Superby (2007) used decision trees, neural networks and linear discriminant analysis for the early identification of three categories of students: low, medium and high-risk students. Some of the background information (demographics and academic history) of the first-year students in Belgian French-speaking universities were significantly related to academic success. Those were: previous education, number of hours of mathematics, financial independence, and age, while gender, parent's education and occupation, and marital status were not significantly related to the academic success. However, all three methods used to predict academic success did not perform well. Overall the correct classification rate was 40.63% using decision trees, 51.88% using neural networks and the best result was obtained with discriminant analysis with overall classification accuracy of 57.35%.

Yu et al. (2007) used a data mining approach to differentiate the predictors of retention among freshmen enrolled at Arizona State University. Using the classification tree based on an entropy tree-splitting criterion they concluded that 'cumulated earned hours', i.e. credits, was the most important factor contributing to retention. Gender and ethnic origin were not identified as significant.

Al-Radaideh, Al-Shawakfa & Al-Najjar (2006) used classification trees to predict the final grades among undergraduate students of the Information Technology & Computer Science Faculty, at Yarmouk University in Jordan. High school grade contributed the most to the separation of students in different clusters. Among background variables gender (both students and lecturers), place of residence, and funding were used to grow the classification tree. However, the classification accuracy was very low, about 35% on average.

Cortez & Silva (2008) predicted the secondary student grades of two core classes using past school grades, demographics, social and other school related data. The results were obtained using data mining techniques such as decision trees, random forests, neural networks and support vector machines. They achieved high level of predictive accuracy when the past grades were included. In some cases their models included also the school related features, demographics (student's age, parent's job and education) and social variables. Unfortunately most of their variables (e.g. student previous grades) were not available for Open Polytechnic students.

Boero, Laureti & Naylor (2005) found that gender is one of the principal determinants of the probability of dropping out. In the binomial probit model they used, males have a higher probability of dropping out relative to the reference group of females. They also found that increasing age has a significant positive effect. The variable was entered in a quadratic form to allow the effect of age to have a diminishing effect on the dropout probability. With regard to pre- university educational qualifications, the type of school attended had a significant effect on the probability of dropping out.

Herrera (2006) concluded that many variables vary in their success at predicting persistence, depending on the academic level. In other words variables that affect persistence at one academic level won't necessarily affect persistence at a different academic level. This means that different models which differentiate between dropout and persistent students should be constructed for each programme level. The same results could be expected at the course levels. That would mean that we would get different probabilities of leaving or staying on the course even for the same student depending upon the course.

Herrera (2006) also discusses educational resilience, which refers to at risk students who completed a course in a timely manner despite risk factors such as biological or psychosocial factors that increase negative outcomes. She also points to the paradigm shift where the focus is now on success rather than on failure. Identifying factors which contribute to the success of an at risk student might help educational institutions increase students' persistence.

In other data mining studies based on enrolment data the following factors were found to be significant: faculty and nationality (Siraj & Abdoulha, 2009) and the secondary school science mark (Dekker, Pechenizkiy & Vleeshouwers, 2009).

In summary, there is mixed evidence on whether the contribution of background information to the early prediction of student success is significant or not. It depends on the list of variables included, the student population and the classification methods used. Even when the background information was significantly related to the study outcome, the prediction accuracy was pretty low with an overall accuracy of around 60% or less.

2. Research Objectives

The main objective of this study is to explore factors that may impact student study outcomes at the Open Polytechnic, one of the major tertiary education providers in this part of the world specialising in distance education. At the time of enrolment at the Open Polytechnic, the only information, i.e. variables we have about students are those contained in their enrolment forms. The question we are trying to address in this paper is whether we can use the enrolment data alone to predict study outcome for newly enrolled students. This issue has not been extensively examined so far at the Open Polytechnic and this paper attempts to fill the gap. We think that the methodology is applicable to any student population, distance or contact, but the results returned would be different for different populations. More specifically the enrolment data were used to achieve the following objectives:

- Build models for early prediction of study outcomes using student enrolment data
- Evaluate the models using cross-validation and misclassification errors to decide which model outperforms other models in term of classification accuracy
- Present results which can be easily understood by the users (students, academic and administrative staff)

The literature review in the first section identified and discussed determinants of study outcome. The methodology and data section describes the data and the statistical methods and models used in this study. Empirical results are presented in the section that follows. The final section discusses the implications of these results.

3. Framework for Data Mining Process

In this paper we have adopted the data mining definition given in Nisbet, Elder & Miner (2009, p. 17). According to them, data mining is “the use of machine learning algorithms to find faint patterns of relationship between data elements in a large, noisy, and messy data set, which can lead to actions to increased benefit in some form (diagnosis, profit, detection, etc.)”.

The framework for data mining applications is based on the CRISP-DM Model created by a consortium of NCR, SPSS, and Daimler-Benz companies. The modified version of the CRISP-DM model is presented on Figure 1, following the project through the general life cycle from business and data understanding, data preparation, modelling, evaluation and deployment. The feedback from deployment to data and business understanding illustrates the iterative nature of a data mining process.

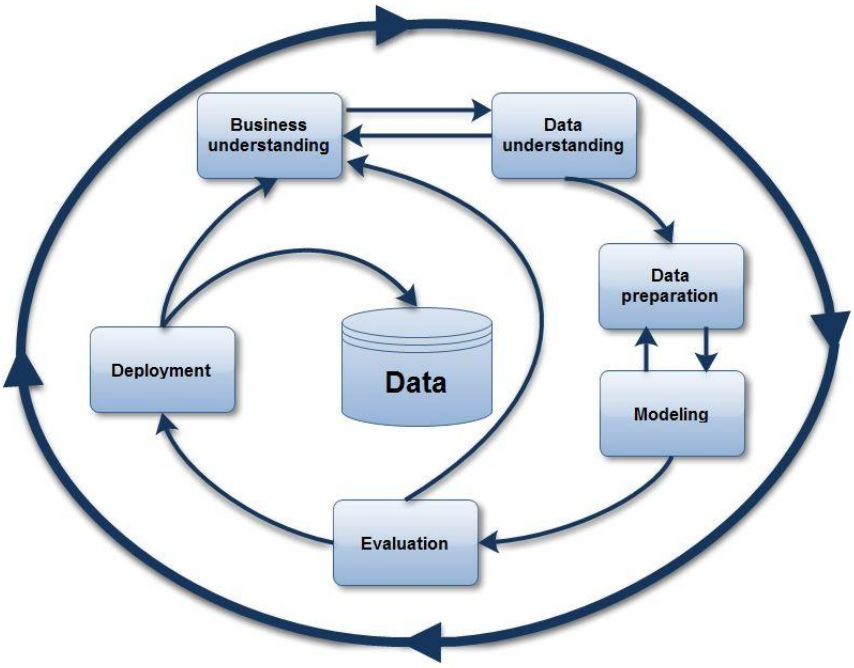


Figure 1: Modified CRISP-DM Model Version 1
(Adopted from Nisbet et al., 2009)

The business understanding phase begins with setting goals for the data mining project. In this paper the goal is an increasing understanding of the pre-enrolment factors that may prevent students from successfully completing the course.

The scope of our research in terms of data used is limited by the data available in the Open Polytechnic Student Management System (known as Integrator) and the enrolment form used for collecting data from newly enrolled students. It is important to have a full understanding of the nature of the data and how it was collected and entered before proceeding further. In this phase an initial data exploration using a pivot table was also conducted to get some insight into the data.

Data preparation is the most important and the most time consuming phase in data mining. Usually 80% of the research time is spent on this phase alone. In this phase the data are put into a form suitable for the modelling phase. If required some selected variables are combined, transformed or used to create new variables. For example, enrolment date and the course block start date were used to generate a variable labelled as "early enrolment". Any data excluded from the data set is documented and their removal explained. Data are cleaned for any duplication of records. For example, in the case of the *Information Systems* course, the course code changed in the past. If a student enrolled during the time when the change in the course code happened and then re-enrolled on the same course, two records exist in the data set for the same student and the same course, but under two different course codes. In this case data for this student were merged into one, single record. The dependent variable "study outcome" with three possible outcomes (labelled as *Pass*, *Fail* and *Lost*) indicates whether students successfully completed the course, failed the course due to not fulfilling course pass requirements or because they voluntary transferred or withdrew or were academically withdrawn from the course. We have defined three different versions of dependent variable (see Table 17)

In the modelling phase we chose and ran models on the training data set. Then we decided whether a suitable model for the data set was found that was acceptable from both an analytical and a managerial standpoint. In this phase we decided to use classification tree models, logistic regression and discriminant analysis. We chose three the most common approaches used by previous studies as described in the literature review. We wanted to compare them for their ability to accurately identify 'at risk' student. The other reasons for such decision are the following: classification tree is a transparent method of classification, easy to apply and interpret its results for both the data miner and the final user of the results. Logistic regression and discriminant analysis are traditionally used in retention studies and we wanted to compare their performance with the classification trees performance.

The evaluation phase involves an iterative process of fitting different versions of models to a training and testing data set, each time evaluating their predictive performance. Once we decided on the final model we can apply it to current data not used during the modelling and evaluation phase.

4. Data and Methodology

The Open Polytechnic student management system does not provide data in a format ready for an easy and direct statistical analysis and modelling. The same problem was reported for the UK Open University (Woodman, 2001). Therefore a data preparation and cleaning as well as the creation of variables for analysis were undertaken to prepare the database for modelling.

4.1 Data preparation

Variables definition and their domains are presented in Table 17. A numeric continuous variable such as age was converted into a categorical variable with only three age groups: under 30, between 30 and 40 and above 40.

Some data mining and multivariate statistical methods are not able to deal with categorical variables measured on a nominal scale, but require a numerical variable. Therefore, for these categorical variables we also created dummy variables, each with two possible values: 1 and 0. For example, variable Māori takes value 1 if the student belongs to NZ Māori ethnic group and 0 otherwise (i.e. belongs to any other ethnic group).

Courses such as: *71238H Business Environment Analysis*, *72395H Environment Economics*, *71251H Information Technology*, *74104H Introduction to Humanities*, *74105H Humanities World Views*, *74106H From Enlightenment to Renaissance*, *74305 Renaissance in Europe*, and *74208 The shape of the world* changed their name during the observed period. For these courses we replaced the old course code and name with the current one.

Other courses, e.g. *71150 Introduction to Information Systems and Technology* changed name and also the offer type. This particular course was first offered as a distance course with online support to be later offered as an online course. We use the same offer type, course code and name of the current course.

From the initial dataset all students granted cross-credit, credit or unspecified credit (course codes: *71297*, *71298*, *71299*, *74198*, and *74199*) were excluded because they didn’t actually study our courses. The courses they had previously completed were recognised and credited to Open Polytechnic courses. The total number of data was reduced to 19468 degree students.

We needed to clarify the definition of categories for the study outcome that we used in our analysis. We considered three possible categories labelled as: *Pass*, *Fail* and *Lost*. Students labelled *Pass* successfully completed the course. Students labelled *Fail* stayed on the course until the end of the course but scored less than the course pass mark. Students labelled *Lost* transferred or withdrew from the course voluntarily or they were withdrawn because they had not completed the in-course assessments.

In data mining variables are also known as features, predictors or attributes. We will use them interchangeably as suggested by Nisbet, Elder & Miner (2009).

4.2 Methodology

Three types of data mining approaches were conducted in this study. The first approach is descriptive which is concerned with the nature of the dataset such as the frequency table and the relationship between the attributes obtained using cross tabulation analysis (contingency tables). In addition, feature selection is conducted to determine the importance of the prediction variables for modelling study outcome. The third type of data mining approach, i.e. predictive data mining is conducted by using two different types of classification trees. We have also estimated the logistic regression models and constructed the discriminant functions for each of three dependent variables. The classification tree models have some advantages over traditional statistical models such as logistic regression and discriminant analysis traditionally used in retention studies. First, they can handle a large number of predictor variables, far more than the logistic regression and discriminant analysis would allow. Secondly, the classification tree models are non-parametric and can capture nonlinear relationships and complex interactions between predictors and dependent variable. We decided not to use other data mining techniques such as neural networks and support vector machines even though in some cases they could achieve higher accuracy, because their structure is not transparent and usually described as a ‘black box’. It is also difficult to explain their results and how they work to a user who would like to apply them to a new set of data.

Finally, a comparison between these models was conducted to determine the best model for the dataset. Data were analysed using SPSS 17 and Statistica 8.

5. Results and Discussion

Before growing the classification trees we summarized the variables by categories and by study outcome, i.e. whether students passed or failed the course. Feature selection was used to rank the variables by their importance for further analysis. Results of the classification tree, estimated logistic regression and discriminant functions are discussed and compared.

5.1 Summary statistics

As part of the data understanding phase we carried out a cross-tabulation for each variable and the study outcome after preparing and cleaning the data. Table 1 reports the results. Based on the results a majority of Open Polytechnic students are female (over 70%). However the percentage of female students who successfully complete their course is slightly higher (70.9%) which suggests that female students are slightly more likely to pass the course than their male counterparts. However, they are also more likely to transfer the course or withdraw from the course than male students.

When it comes to age, over 68% of students are above 30, with the majority in the age group between 30 and 40. This age group is also more likely to fail the course because the percentage of students who failed the course in this age group (39.7%) is higher than their overall participation in the student population (38.6%). Disability was shown to be a disadvantage for Open Polytechnic students. Students with a disability are more likely to fail than those without a disability. There are huge differences in the percentage of students who successfully completed courses depending on their ethnic origin. Though Māori are 5.5% of all students, their participation is significantly lower in the *Pass* subpopulation (i.e. 3.3%) and higher in the *Fail* subpopulation (6.1%). The situation is even worse with Pacific Islands students. They are 3.5% of all students, but their participation is significantly lower in the *Pass* subpopulation (1.9%) and much higher in the *Fail* subpopulation (7.6%). Based on these results we can say that students within these ethnic groups are identified as students 'at risk'. Further methods of data mining, logistic regression and discriminant analysis will confirm this statement.

A substantial number of students (over 40%) don't have a secondary school qualification higher than NCEA Level 2 on the New Zealand National Qualification Framework and are more vulnerable than the other categories in this variable. Over two-thirds of Open Polytechnic students are working and studying at the same time. Though the difference between those who work and those who are not working is not substantial, it is interesting to note that the students who are working are more likely to pass the course than those not working.

We used "Early enrolment" as a proxy for motivation and good time management skills. Students who are motivated and are planning their study in advance will also enrol well before the enrolment closing date. The opposite category "late comers" makes 30% of the total number of students, but these students are more likely to fail the course. Their participation in the *Fail* subpopulation increased from 30% to 32.2%.

Almost one third of students are enrolled on the Bachelor of Applied Sciences programme. They are more likely to fail the course when compared with students enrolled on the Bachelor of Arts programme. Finally, students studying in the summer semester (Semester 3) are more likely to fail than those studying in the first and second semester. The reasons might be that this semester has an increased number of transferees due to the overlap between Semester 2 and Semester 3 and because there are increased distractions in the summer (Christmas and summer holidays).

Table 1: Descriptive statistics (percentage) – Study outcome 1 (19468 students)

Variable	Domain	Count	Total	Pass	Fail	Lost
Gender	Female	13744	70.6	70.9	66.4	73.2
	Male	5724	29.4	29.1	33.6	26.8
Age	<30	4879	25.1	21.7	35.5	24.7
	30-40	6882	35.4	34.9	35.5	36.3
	>40	7707	39.6	43.4	29.0	39.0
Disability	Yes	1405	7.2	6.3	7.6	9.0
	No	18063	92.8	93.7	92.4	91.0
Ethnicity	European	644	3.3	3.8	2.2	2.9
	Chinese	588	3.0	3.3	2.8	2.6
	Pakeha	14131	72.6	76.4	62.8	71.5
	Asian	479	2.5	2.5	2.6	2.2
	Others	1093	5.6	5.6	5.1	6.2
	Indian	777	4.0	3.3	6.1	4.0
	Māori	1067	5.5	3.3	10.8	6.5
	Pacific	689	3.5	1.9	7.6	4.2
Secondary school	No	1529	7.9	5.5	13.4	9.1
	NCEA Level 1	2511	12.9	12.0	14.3	14.0
	NCEA Level 2	3843	19.7	19.3	20.0	20.6
	University Entrance	4645	23.9	25.4	20.1	23.2
	NCEA Level 3	2628	13.5	15.0	11.0	11.9
	Overseas qualification	3478	17.9	19.3	15.5	16.4
	Other	834	4.3	3.5	5.8	4.8
Work status	Working	13549	69.6	72.0	65.1	67.5
	Not working	5919	30.4	28.0	34.9	32.5
Early enrolment	Yes	15405	79.1	80.4	75.1	79.3
	No	4063	20.9	19.6	24.9	20.7
Course faculty	School of Business	8687	44.6	44.1	48.0	43.2
	School of I&S Science	10024	51.5	51.7	47.7	54.0
	Workplace Learning	757	3.9	4.2	4.4	2.8
Course programme	Bachelor of Business	11247	57.8	57.1	59.5	57.9
	Bachelor of Appl. Sci.	5987	30.8	29.5	32.0	32.7
	Bachelor of Arts	2234	11.5	13.4	8.5	9.4
Course level	Level 5	10229	52.5	45.0	66.0	59.5
	Level 6	5394	27.7	32.3	18.7	24.1
	Level 7	3845	19.8	22.7	15.4	16.4
Course block	First	8269	42.5	44.9	38.4	40.1
	Second	8719	44.8	44.2	44.2	46.6
	Third	2480	12.7	10.9	17.4	13.4
Course offer type	Distance	15117	77.7	79.8	73.2	76.3
	Blended	1833	9.4	8.9	10.3	10.0
	Open	2518	12.9	11.4	16.6	13.7

Each of the variables used in this research study are also graphically presented in Appendix C as well as with the Study outcome 1 in Table 1. We can say that the successful course completion increases with age and also with the level of the secondary school qualification. The highest successful course completion occurred in Semester 1 (60%) and then decreases in Semester 2 (56%) and even more in Semester 3 (48%) due to the factors explained in the previous paragraph.

5.2 Feature selection

The number of predictor variables is not so large and so we don't have to select a subset of variables for further analysis which is the main purpose of applying feature selection to data. However, feature selection could be also used as a pre-processor for predictive data mining to rank predictors according to the strength of their relationship to dependent or outcome variables. During the feature selection process no specific form of relationship, neither linear nor nonlinear is assumed. The outcome of the feature selection would be a rank list of predictors according to their importance for further analysis of the dependent variable with the other methods for regression and classification.

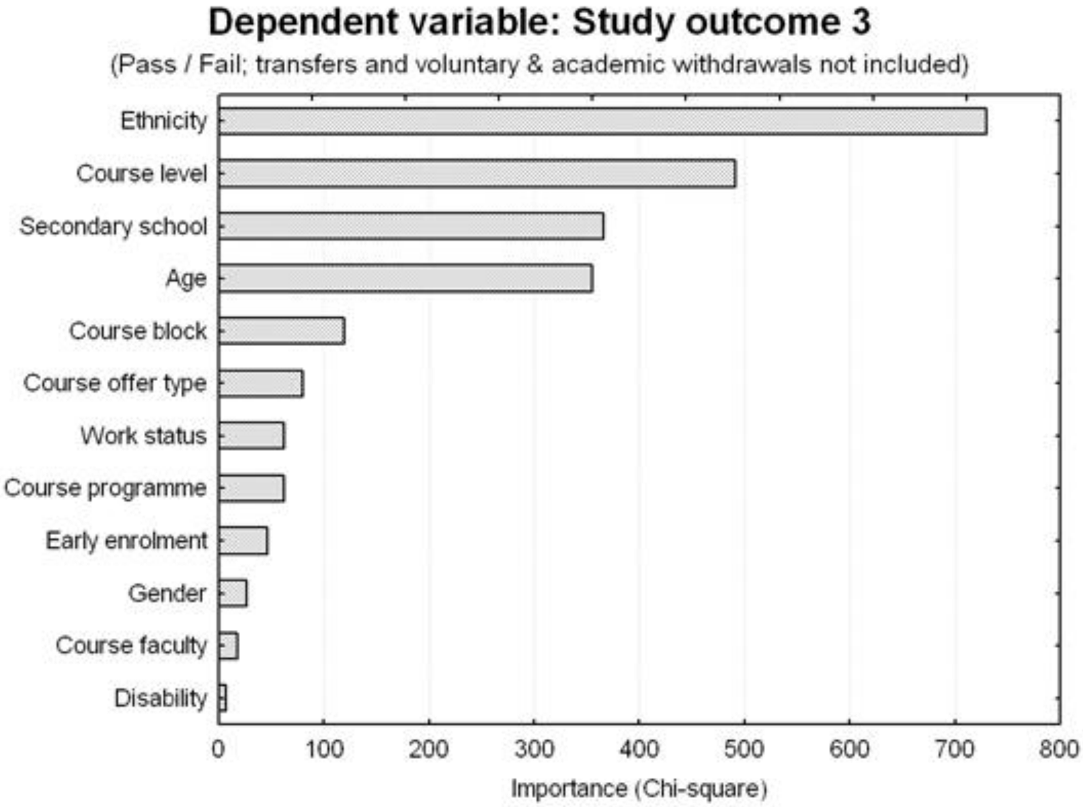


Figure 2: Importance plot for predictors (Study outcome 3)

The results of feature selection are presented in Figure 2 and also in Table 2. Figure 2 shows the importance plot for Study outcome 3, i.e. the dependent variable where the *Lost* students were excluded from the data set. The chi-square statistic in Table 2 for Study outcome 3 is a measure of how important a particular feature is for a study outcome. The smaller the *P*-value of the chi-square test, the stronger the evidence that a particular feature is important. It shows that all features are statistically significant. The features are sorted in decreasing order, i.e. from the most to the least important.

To decide how many features to select for further analysis we have two options. We can either select the top 4 (significantly higher Chi-square values than the rest of the variables), or we can look for other inflection points in the curve and select the top 6 or even top 8 because after the top 8 variables, the remainder level off in a plateau effect.

Table 2: Best predictors for dependent variable (Study outcome 3)

Variable	Chi-square	*P*-value
Ethnicity	730.19	0.00
Course level	491.82	0.00
Secondary school	365.67	0.00
Age	355.60	0.00
Course block	119.78	0.00
Course offer type	80.21	0.00
Work status	62.63	0.00
Course programme	62.57	0.00
Early enrolment	46.79	0.00
Gender	27.18	0.00
Course faculty	18.01	0.00
Disability	7.30	0.01

In all three cases, i.e. for all three definitions of the dependent variable, if the top 8 variables are selected, we get the same list of predictors. Therefore we can conclude that the list of important predictors is quite robust to changes in the study outcome definition. We may proceed into the next step using the top 8 variables:

1. Ethnicity
2. Course level
3. Secondary school
4. Age
5. Course block
6. Course programme
7. Course offer type
8. Work status

Though the results of the feature selection might suggest continuing analysis with only the subset of predictors, we have included all available predictors in our classification tree analysis. We followed the advice given in Luan & Zhao (2006) who suggested that even though some variables may have little significance to the overall prediction outcome, they can be essential to a specific record in our data set. For example, for Indian students only the ‘early enrolment’ variable contributed significantly to the separation of successful from unsuccessful students. Therefore we kept the ‘early enrolment’ variable in the model even though it was important only for a very specific subpopulation of students.

5.3 Comparison of different models

In this section we summarise and compare the accuracy of all estimated models. A detailed discussion of individual estimated models is given in the following sections. Generally there are a few advantages that classification tree models have over logistic regression and discriminant analysis when applied to student enrolment data and student retention phenomenon.

First, classification trees, logistic regressions and discriminant analysis achieved almost the same predictive accuracy as measured by the overall percentage of correct classification (with one exception explained below). However, in the case of logistic regressions and discriminant analysis that was achieved at the cost of including between 3 and 8 times more variables. This is due to the way the classification trees are constructed. Namely, the classification tree is not using all the statistically significant predictors at the same time as the logistic regression and discriminant analysis, but it uses only those predictors, one by one, which contribute to the best split at each stage. Second, while the logistic regression and discriminant analysis generally neglect interactions between covariates, assuming all covariates are independent, the classification tree takes this interaction into account. If the interactions between covariates are significant, which is a reasonable assumption for independent variables in the logistic regressions, then the odds ratios are not quite appropriate measures of the impact that an independent variable could have on a dependent variable. Third, the logistic regression and discriminant analysis separate the students into two groups: *Pass* and *Fail*. However, the classification tree classifies students into more than two groups (e.g. 18 groups in case of CHAID model and Study outcome 3) providing additional information about successful and unsuccessful students, i.e. a more detailed description of their profiles. Finally, the classification tree approach is simple to use. Its results are easy to interpret and apply to a newly enrolled student. By asking a few questions as described in the tables with classification rules, a student can be classified as *Pass* or *Fail* with a probability allocated to each profile. There is no need to use odds ratios to calculate the scores and probability for each individual or discriminant function scores or to compare them with the threshold value to see whether a student should be classified as an 'at risk' student. Taking all these into consideration we would suggest the use of the classification tree in retention studies for classifying and describing 'at risk' students.

Table 3 summarises the classification accuracy of the four estimated models and gives the number of variables used to achieve it. Logistic regression and discriminant analysis achieved almost the same level of accuracy for all three study outcome variables. However, the number of variables used in these models was significantly higher than the number of variables used in classification tree models.

For Study outcome 1 the difference in accuracy between the classification tree models (CHAID and CART) and the other two models (logistic regression and discriminant analysis) is due to the use of different misclassification costs. We assigned double cost to the classification outcome that predicts a student will pass when in fact the student failed the course (details in Section 5.4). Though the overall accuracy of the classification tree models decreased, they perform better at more accurately identifying 'at risk' students that would be classified as successful, i.e. pass, than the other two models. For example, from Table 23 students who failed the course would be correctly identified in 44.1% cases and those who transferred or withdrew in 40.7% cases. For the same study outcome variable discriminant analysis correctly identifies students who failed the course in 22.7% cases and those who transferred or withdrew in only 1.9% cases. So the price paid for higher accuracy in identification of ‘at risk’ students was a decrease of the overall accuracy.

For the second definition of the study outcome all four models achieved almost the same level of accuracy. However, the CART model requires only four variables while discriminant analysis needs almost 8 times more variables to achieve the same accuracy.

Table 3: Comparison of different models (accuracy and number of variables)

Study outcome	Model	Accuracy	Number of variables
Study outcome 1: *Pass*, *Fail* & *Lost* (transfers & withdrawals – academic and voluntary)	CHAID	46.5%	7
	CART	47.1%	8
	Logistic regression	58.3%	21
	Discriminant analysis	58.3%	32
Study outcome 2: *Pass* & *Fail* (includes *Lost*, i.e. transfers & withdrawals – academic and voluntary)	CHAID	62.4%	6
	CART	63.0%	4
	Logistic regression	63.9%	27
	Discriminant analysis	64.0%	31
Study outcome 3: *Pass* & *Fail* (excludes *Lost*, i.e. transfers & withdrawals – academic and voluntary)	CHAID	73.1%	7
	CART	75.2%	5
	Logistic regression	77.0%	29
	Discriminant analysis	77.0%	30

For the third definition of the study outcome logistic regression and discriminant analysis slightly outperform the classification tree models. The CART model is more accurate than the CHAID model and is the most parsimonious model among all four models using only five variables to achieve 75.2% overall accuracy. The CART model achieves a relatively high accuracy level (comparable to accuracy of other models) with the smaller number of variables and therefore we would recommend its use for early identification of ‘at risk’ students.

5.4 Classification trees

The classification tree recursively partitions the data into two or more groups that are more homogeneous in the following steps. The resulting classification rules are contained in the path from the initial, i.e. root node to the terminal node or leaf. As discussed in Nisbet, Elder & Miner (2009, p. 140) there are three elements that define a classification tree algorithm:

1. For each node a specific rule describes splitting the data on one variable
2. A stopping rule is defined to decide when to stop growing the tree further
3. Each terminal node is assigned to an outcome, i.e. the prediction of the dependent variable

The objective of an analysis based on a classification tree is to identify factors that contribute the most to the separation of successful from unsuccessful students. When the classification tree is formed we can calculate the probability of each student being successful. Once the classification tree is formed, it could be used in the new data set to predict the study outcome for newly enrolled students. Details about criteria and the procedure for merging classes and selecting the split variable and the stopping criteria are explained and discussed in detail in Hastie, Tibshirani & Friedman (2009), Han & Kamber (2006), Nisbet, Elder & Miner (2009) and Rokach & Maimon (2008).

To evaluate the classification tree model we used part of the data set for training the tree. Once the classification tree model is estimated we are using the same model, but this time with the rest of the data previously not used during the training phase. For each classification tree we have randomly split the data set into training and testing parts with 75% of data used during training and 25% of data used in testing, i.e. evaluation phase. We used two stopping criteria in the training process:

1. A minimum number of cases included in the split has been reached: 300 cases in parent node and 100 cases in the child node
2. A maximum tree depth has been reached: 3 levels for the CHAID tree and 5 levels for the CART tree

Finally for each classification tree we have assigned different costs to the classification outcomes (see the misclassification costs matrix in Table 4). This is one of the options of increasing the percentage of correctly classified unsuccessful students. Since the main objective in the student retention analysis is to build a model that correctly identifies ‘at risk’ students, we assigned double cost to that particular outcome. In other words we penalised the outcome of a model that predicts *Pass*, when in fact the student failed the course.

Table 4: Misclassification costs
(Study outcome 3)

	Predicted	
Observed	Fail	Pass
Fail	0	2
Pass	1	0

5.4.1 CHAID

The acronym CHAID stands for Chi-square Automatic Interaction Detector. A CHAID tree allows for more than two splits to occur from a single parent node (for details see Nisbet, Elder & Miner, 2009). We started our classification tree analysis by growing the tree with equal costs for each outcome and splitting the data set in proportion 75%:25% between training and test data. Once the tree is trained it could be used outside the sample, i.e. in the test data set to predict study outcome for students included in the test data. If we achieve similar accuracy with trees built on both training and test data we can safely use the model predicting outcome with a new data set. Therefore we began our analysis by comparing the accuracy of the same model based on training and test data.

For the model with equal costs we got the following classification matrix (Table 5). Though the overall accuracy of the model using the training data is relatively high (76%) and testing the model produced almost the same accuracy, closer inspection of the other accuracy measures in Table 5 suggest a poor performance of the model. It predicts failure for only 16.2% of unsuccessful students, which means that 83.8% of unsuccessful students are inaccurately classified as successful students. The practical consequence of this misclassification is that these students would not have received the additional learning support provided to the students ‘at risk’, simply because they will be classified among successful students by the model. This feature of the model is more critical than the misclassification of the successful students among unsuccessful students (29% of successful students belong to this category in case of training data).

In this case these students may receive additional learning support or counselling with regard to course choice even though they don’t need it.

Table 5: CHAID classification matrix for training & testing data with equal costs (Study outcome 3)

		Predicted		
	Observed	Fail	Pass	Percent correct
Training	Fail	455	2359	16.2%
	Pass	285	7929	96.5%
	Overall percentage	6.7%	93.3%	76.0%
Test	Fail	147	747	16.4%
	Pass	97	2685	96.5%
	Overall percentage	6.6%	93.4%	77.0%

One option to increase the percentage of correctly classified unsuccessful students is to change the misclassification cost matrix. With this option there is always a trade-off between increasing the percentage of correct classification of unsuccessful students and decreasing percentage of correct classification for successful students as well as decreasing the percentage of overall correct classification.

To illustrate the impact of misclassification costs matrix has on the accuracy result we used both training and testing data with the CHAID tree. This time we used the misclassification costs from Table 4. In this case the increased cost for misclassification of unsuccessful to the successful group of students increased the percentage of correctly classified unsuccessful students from 16.2% (Table 5) to 43.6% (Table 6). This significant improvement of the model accuracy was paid with a small decrease of the overall accuracy from 76% (Table 5) to 73.5% (Table 6).

Table 6: CHAID classification matrix for training & testing data (Study outcome 3)

		Predicted		
	Observed	Fail	Pass	Percent correct
Training	Fail	1196	1548	43.6%
	Pass	1363	6889	83.5%
	Overall percentage	23.3%	76.7%	73.5%
Test	Fail	416	548	43.2%
	Pass	464	2280	83.1%
	Overall percentage	23.7%	76.3%	72.7%

The estimates of the risk presented in Table 7 are 0.406 and 0.421 for the training and test data respectively. They indicate that the category predicted by the model (successful or unsuccessful student) is wrong for 41% and 42% of the cases respectively. So the risks of misclassifying a student are approximately 41% and 42%. These results are quite consistent with the results in the CHAID classification matrix (Table 6) where these percentages are 43.6% and 43.2% respectively. The differences in these percentages are due to the different costs assigned to outcomes. As we said, since the main objective is to build a model that correctly identifies ‘at risk’ students, we assigned two times higher cost to that particular outcome. In other words we penalised the outcome of the model that predicts *Pass*, when in fact the student failed the course.

Table 7: Risk for CHAID model
(Study outcome 3)

Sample	Estimate	Standard Error
Training	0.406	0.007
Test	0.421	0.012

The classification tree grown with the test data have achieved almost the same accuracy as the tree grown with the training data. That would suggest that the model performed well. Therefore we decided to combine both training and test data into one data set and re-grow the trees with a complete data set. We also grew the tree using misclassification costs greater than 2 (results are not presented), but the classification accuracy of these trees dropped significantly.

The rectangles in Figure 3 represent a node in the classification tree. Each node contains the following information: the number of successful students (4th line, last column) and unsuccessful students (3rd line, last column), as well as the percentages for each category (2nd column) and the relative and absolute size of the node (5th line). The variable names above the nodes are the predictors that provided the best split for the node according to the classification and regression tree-style exhaustive search for univariate splits method. This method looks at all possible splits for each predictor variable at each node. The search stops when the split with the largest improvement in goodness of fit, based on the Gini measure of node impurity (for CART), is found. Immediately above the nodes are categories which describe these nodes. Note that all available predictor variables in the dataset were included in the classification tree analysis in spite of their insignificance as detected in the feature selection section.

The CHAID classification tree generated the tree structure presented in Figure 3 (Māori & Pacific Islander branch), while the other branches are presented in Appendix C. It shows that the following variables were used to construct the tree: (1) ethnicity, (2) course level, (3) course programme, (4) course faculty, (5) age, (6) gender, (7) secondary school and (8) early enrolment. All the other variables were used but not included in the final model. We could change the stopping criteria to allow further growing of the tree. That would probably allow other variables to enter the model, but that would also result in nodes with just a few students. In the most extreme case we can continue splitting the tree until we create a terminal node for every student. However, we would get a model, i.e. classification tree that fits data better, but with a likely poorer performance when used on a new data set. This phenomenon is known as overfitting.

The largest successful group (i.e. students who successfully completed the course) consists of 3455 (23.5%) students (Node 17). The ethnic origin of students in this group is either Pakeha or Chinese. Students in this group studied Level 6 or 7 courses in the School of Information and Social Sciences or Workplace Learning and Development. The largest unsuccessful group (i.e. students who were unsuccessful) contains 533 students (3.6% of all students) (Node 26). They are either Māori or Pacific Island students studying Level 5 courses toward Bachelor of Business or Bachelor of Arts. The next largest group considered also as unsuccessful students, contains 249, i.e. 1.7% of all students, where 75.1% of them are unsuccessful (Node

25). They are described as Māori or Pacific Islands students studying Level 5 courses toward Bachelor of Applied Science.

The overall percentage of correct classification for the study outcome is 73.1% (Table 8). This percentage of correct classification was achieved using only 8 variables.

Table 8: CHAID classification matrix
(Study outcome 3)

	Predicted		
Observed	Fail	Pass	Percent correct
Fail	1638	2070	44.2%
Pass	1883	9113	82.9%
Overall percentage	23.9%	76.1%	73.1%

The cross-validation estimate of the risk is 0.41 and indicates that the category predicted by the model (successful or unsuccessful student) is wrong for 41% of the cases. So the risk of misclassifying a student is approximately 41%. This result is not quite consistent with the results in the CHAID classification matrix (Table 8) because of different costs assigned to outcomes. Since the main objective is to build a model that correctly identifies ‘at risk’ students, we assigned a cost two times higher to that particular outcome. In other words we penalised the outcome of the model that predicts *Pass*, when in fact the student failed the course.

With numbers of false positives (2070) and false negatives (1883), the CHAID tree is in itself still not reasonably accurate at identifying an unsuccessful student (positive predictive value is 44.2%) though we increased the cost for this particular outcome. It will pick up only 23.9% of all unsuccessful students (known as the sensitivity). The predictive values, which take into account the prevalence of failing the course, are generally more important in determining the usefulness of a prediction model. The negative predictive value was of more concern to the course because the objective was to minimize the probability of being in error when deciding that a student is not at risk for not completing the course. However the CHAID model, as a classification tool, will pick-up with high probability successful students (negative predictive value is 82.9%) and correctly identifies 76.1% of those who pass the course (known as the specificity).

The classification matrix also indicates another problem with the model. It predicts failure for only 44.2% of unsuccessful students, which means that 55.8% of unsuccessful students are inaccurately classified with the successful students. The practical consequence of this misclassification is that these students would not receive additional learning support provided to the students ‘at risk’, simply because they will be classified among successful students by the model. This feature of the model is more critical than the misclassification of the successful students among unsuccessful students (17.1% of successful students belong to this category). Because in this case these students may receive additional learning support even though they don’t need it. As we said, one option to increase percentage of correctly classified unsuccessful students is to change the misclassification cost matrix as we have done.

Another tool used to assess the quality of the model is the gains chart (known also as a lift chart). For the CHAID classification tree the gains chart is presented in Figure 4. For a good model the gains chart will rise steeply toward 100% and then will curve down. The gains chart line close to the diagonal reference line indicates that the model does not work well, i.e. not separating well successful from unsuccessful students.

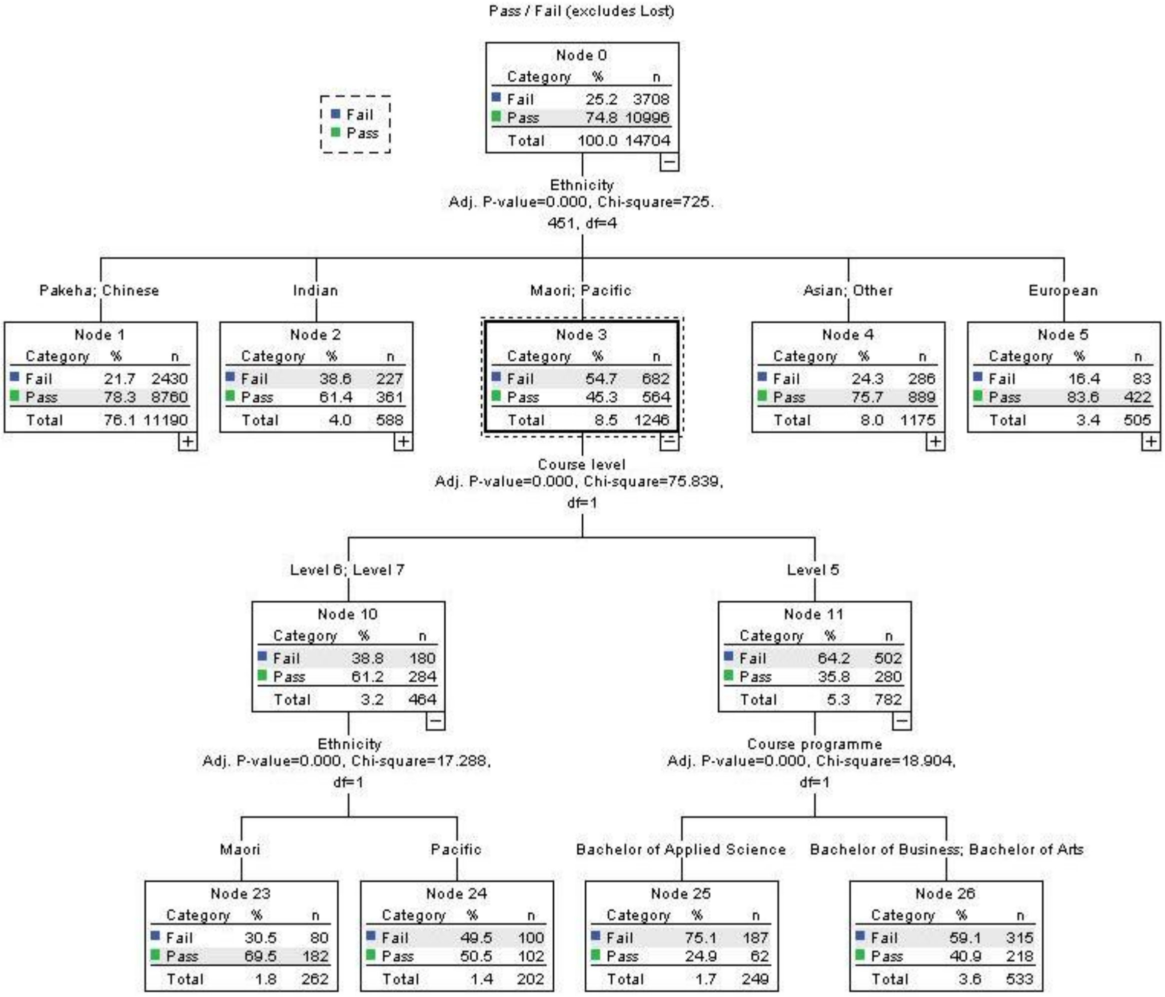


Figure 3: CHAID tree (Study outcome 3: Māori & Pacific Islanders)

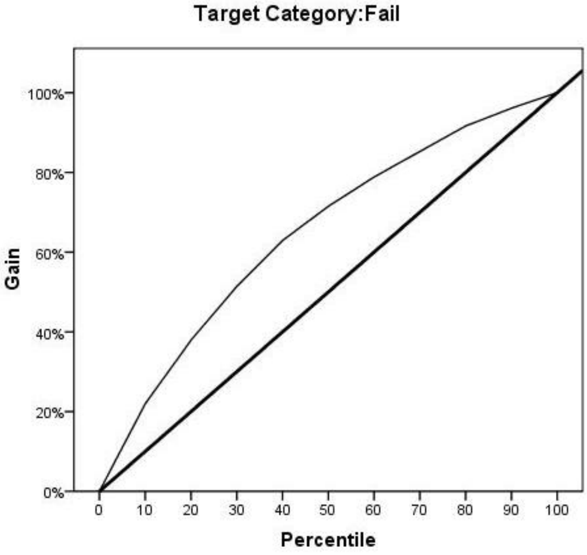


Figure 4: Gain chart for unsuccessful student (*Fail* category) - CHAID

Note that the classification accuracy of our model is about 76%. This could indicate that we may need to do more work (either in preprocessing or in selecting the correct parameters for classification), before building another model.

Once the tree is grown we can write down the classification rules by simply following the tree from the initial node to each terminal node. Rules can be used for a simple explanation of the results and also for deciding on the study outcome for the newly enrolled student. They can be written in IF-THEN format. Rules for the CHAID tree (Study outcome 3: Māori & Pacific Islanders branch) for all four terminal nodes are given in Table 9.

The CHAID classification tree in Node 24 does not make a clear distinction between successful and unsuccessful students, because the probabilities of passing (0.505) or failing the course (0.495) for students in this node are almost equal. The "Māori & Pacific Islanders branch" in the CHAID classification tree in Figure 3 suggests that Māori and Pacific Islands students need additional learning support to increase their chance of successful completing the course.

Table 9: Rules for CHAID tree (Study outcome 3: Māori & Pacific Islanders)

Node	Rule	Outcome	Probability
23	**IF** Ethnicity = "Māori" **OR** "Pacific Islanders" **AND** Course level = "Level 6" **OR** "Level 7" **AND** Ethnicity = "Māori" **THEN**	Pass	0.695
24	**IF** Ethnicity = "Māori" **OR** "Pacific Islanders" **AND** Course level = "Level 6" **OR** "Level 7" **AND** Ethnicity = "Pacific Islanders" **THEN**	Pass	0.505
25	**IF** Ethnicity = "Māori" **OR** "Pacific Islanders" **AND** Course level = "Level 5" **AND** Course programme = "Bachelor of Applied Science" **THEN**	Fail	0.751
26	**IF** Ethnicity = "Māori" **OR** "Pacific Islanders" **AND** Course level = "Level 5" **AND** Course programme = "Bachelor of Business" **OR** "Bachelor of Arts" **THEN**	Fail	0.591

For other terminal nodes (see Figure 19 for Pakeha & Chinese students, Figure 20 for Indian students, Figure 21 for Asian & Others students and Figure 22 for European students) the probabilities of successfully completing courses are higher than the probabilities of an unfavourable study outcome.

5.4.2 CART

The acronym CART stands for Classification and Regression Tree. While CHAID allows for a multiway split, CART splits the data at each level into only two nodes. We used the standard practice of overgrowing the tree and then we pruned it back to the optimal size. Figure 5 shows the CART classification tree for Study outcome 3. It shows that only five variables were used to construct the tree: (1) ethnicity, (2) course level, (3) age, (4) secondary school and (5) course faculty.

The largest successful group (i.e. students who successfully completed the course) consists of 6485 (44.1%) students (Node 5). The ethnic origin of students in this group is either Pakeha, Asian, Chinese, European or Others. Students in this group were studying Level 6 & 7 courses. Student in this node would pass the course with high probability (0.854). This is because most of these students have probably already passed level 5 courses and their chances of getting more successful scores are increased.

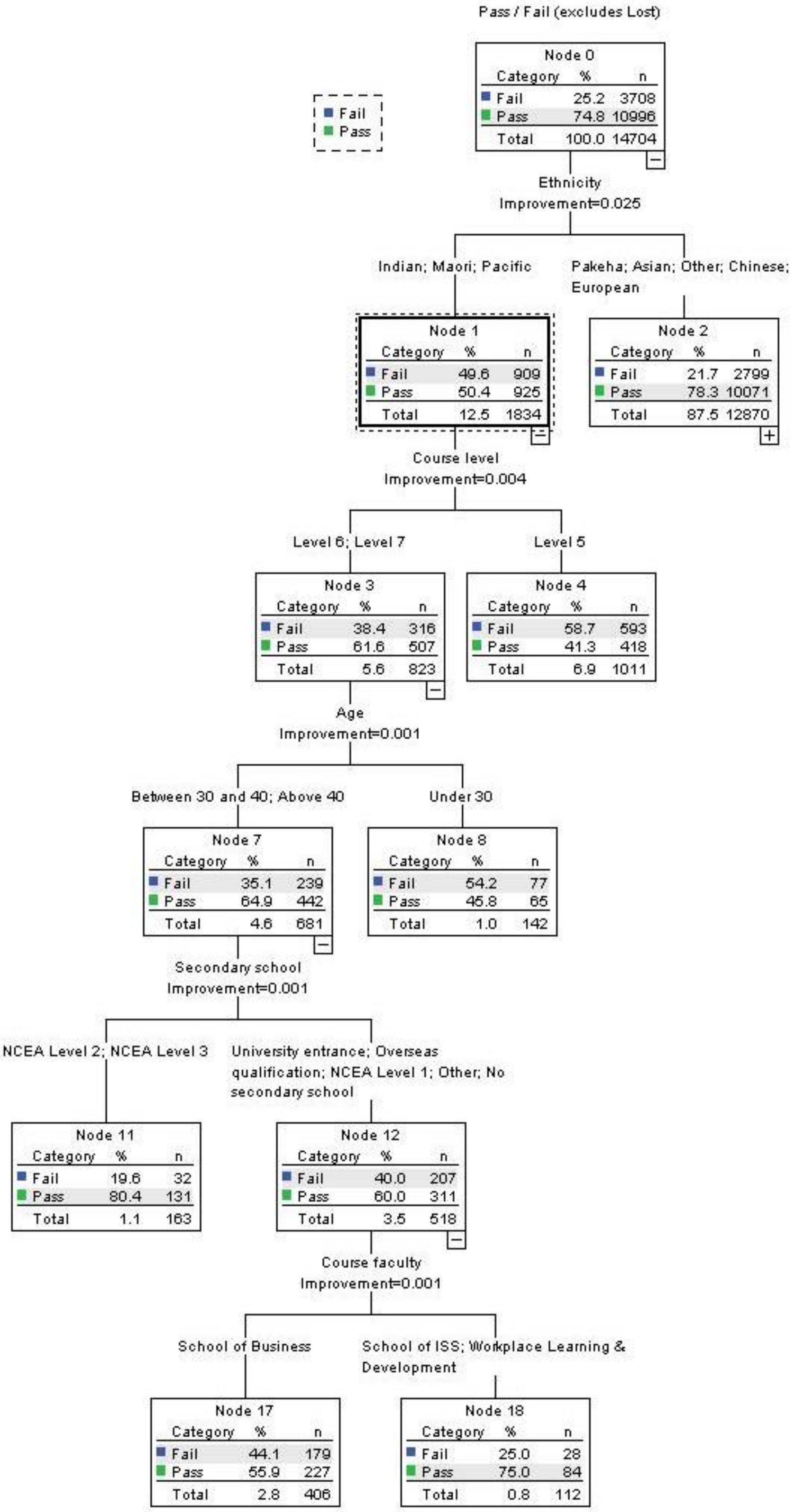


Figure 5: CART tree (Study outcome 3: Indian, Māori & Pacific Islanders)

The largest unsuccessful group (i.e. students who were unsuccessful) contains 1011 students (6.9% of all students) that belong to Node 4. They are Māori, Pacific Islands or Indian students. The accuracy of the classification tree is presented in Table 10.

Table 10: CART classification matrix (Study outcome 3)

	Predicted		
Observed	Fail	Pass	Percent correct
Fail	1524	2184	41.1%
Pass	1463	9533	86.7%
Overall percentage	20.3%	79.7%	75.2%

While the overall accuracy (75.2%) is higher than in CHART model (73.1%) the CART model is less successful at identifying an unsuccessful student; positive predictive value is 41.1% (44.2% with CHAID).

The rules for the CART classification tree (Indian, Māori and Pacific Islands students) are given in Table 11 for all five terminal nodes in Figure 5.

Table 11: Rules for CART tree (Study outcome 3: Māori & Pacific Islanders)

Node	Rule	Outcome	Probability
4	**IF** Ethnicity = "Māori" **OR** "Pacific Islanders" **OR** "Indian" **AND** Course level = "Level 5" **THEN**	Fail	0.587
8	**IF** Ethnicity = "Māori" **OR** "Pacific Islanders" **OR** "Indian" **AND** Course level = "Level 6" **OR** "Level 7" **AND** Age = "Under 30" **THEN**	Fail	0.542
11	**IF** Ethnicity = "Māori" **OR** "Pacific Islanders" **OR** "Indian" **AND** Course level = "Level 6" **OR** "Level 7" **AND** Age = "Between 30 and 40" **OR** "Above 40" **AND** Secondary school = "NCEA Level 2" **OR** "NCEA Level 3" **THEN**	Pass	0.804
17	**IF** Ethnicity = "Māori" **OR** "Pacific Islanders" **OR** "Indian" **AND** Course level = "Level 6" **OR** "Level 7" **AND** Age = "Between 30 and 40" **OR** "Above 40" **AND** Secondary school = "No secondary school" **OR** "NCEA Level 1" **OR** "University entrance" **OR** "Overseas qualification" **OR** "Other" **AND** Course faculty = "School of Business" **THEN**	Pass	0.559
18	**IF** Ethnicity = "Māori" **OR** "Pacific Islanders" **OR** "Indian" **AND** Course level = "Level 6" **OR** "Level 7" **AND** Age = "Between 30 and 40" **OR** "Above 40" **AND** Secondary school = "No secondary school" **OR** "NCEA Level 1" **OR** "University entrance" **OR** "Overseas qualification" **OR** "Other" **AND** Course faculty = "School of Information and Social Sciences" **OR** "Workplace Learning and Development" **THEN**	Pass	0.750

The cross-validation estimate of the risk is 0.397 indicating that the category predicted by the model (successful or unsuccessful student) is wrong for 39.7% of the cases. The CART classification matrix (Table 10) shows that model correctly classifies 75.2% of students. This is a slight increase in comparison to the CHAID model. The numbers of false positives (2184) for the CART model increases, therefore decreasing the positive predictive value to 20.3%. In other words the CHAID model will work better than the CART model at identifying an unsuccessful student. The price paid for increasing the accuracy of the CART model is reflected in decreasing sensitivity. The CHAID model will pick up 23.9% of all unsuccessful students (CART model only 20.3%). At the same time the specificity will increase to 79.7% (CHAID model 76.1%).

Figure 6 shows the second branch of the CART tree, i.e. Pakeha, Chinese, European, Asian and Others students.

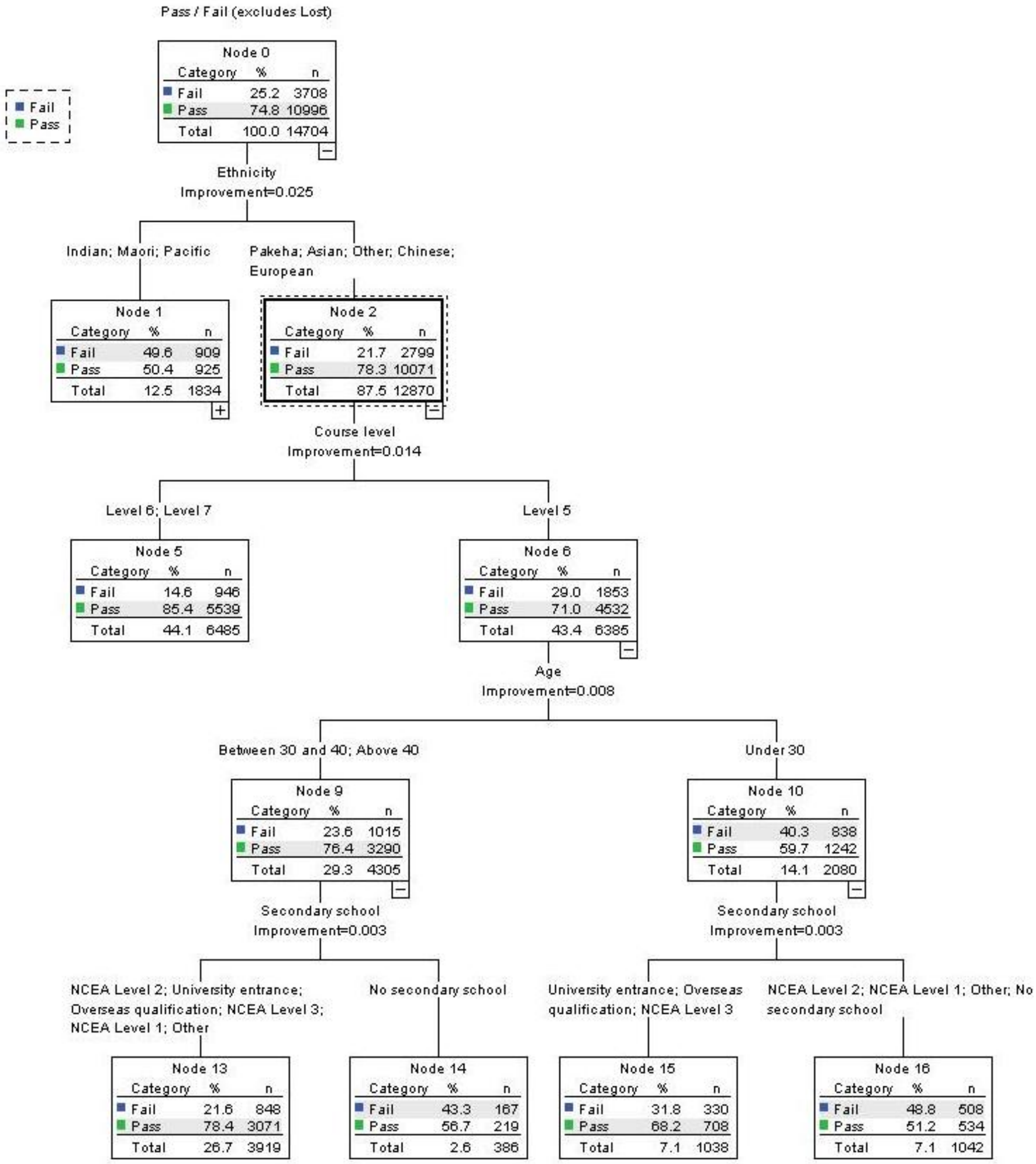


Figure 6: CART tree (Study outcome 3: Pakeha & Asian)

The rules for the CART classification tree are given in Table 12 for all five terminal nodes in Figure 6. These rules could be used with a new data set to decide on the possible study outcome for a newly enrolled student.

Table 12: Rules for CART tree (Study outcome 3: Pakeha & Asian)

Node	Rule	Outcome	Probability
5	**IF** Ethnicity = "Pakeha" **OR** "Asian" **OR** "Chinese" **OR** "European" **OR** "Others" **AND** Course level = "Level 6" **OR** "Level 7" **THEN**	Pass	0.854
13	**IF** Ethnicity = "Pakeha" **OR** "Asian" **OR** "Chinese" **OR** "European" **OR** "Others" **AND** Course level = "Level 5" **AND** Age = "Between 30 and 40" **OR** "Above 40" **AND** Secondary school = "NCEA Level 1" **OR** "NCEA Level 2" **OR** "NCEA Level 3" **OR** "University entrance" **OR** "Overseas qualification" **OR** "Other" **THEN**	Pass	0.784
14	**IF** Ethnicity = "Pakeha" **OR** "Asian" **OR** "Chinese" **OR** "European" **OR** "Others" **AND** Course level = "Level 5" **AND** Age = "Between 30 and 40" **OR** "Above 40" **AND** Secondary school = "No secondary school" **THEN**	Pass	0.567
15	**IF** Ethnicity = "Pakeha" **OR** "Asian" **OR** "Chinese" **OR** "European" **OR** "Others" **AND** Course level = "Level 5" **AND** Age = "Under 30" **AND** Secondary school = "NCEA Level 3" **OR** "University entrance" **OR** "Overseas qualification" **THEN**	Pass	0.682
16	**IF** Ethnicity = "Pakeha" **OR** "Asian" **OR** "Chinese" **OR** "European" **OR** "Others" **AND** Course level = "Level 5" **AND** Age = "Under 30" **AND** Secondary school = No secondary school" **OR** "NCEA Level 1" **OR** "NCEA Level 2" **OR** "Other" **THEN**	Pass	0.512

The gains chart for the CART classification tree is presented in Figure 7. Gains charts for two models are almost the same, with slightly larger gains obtained with the CART model.

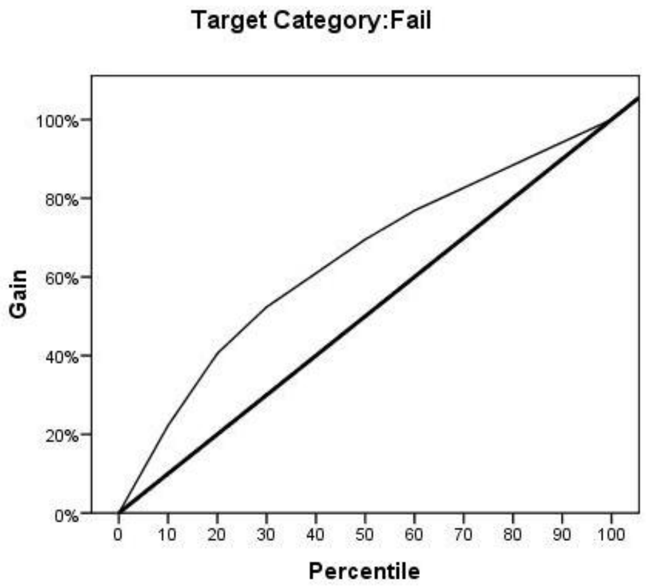


Figure 7: Gain chart for unsuccessful student (*Fail* category) - CART

The classification tree results for the study outcome suggest that among the student demographics information such as gender, age, ethnicity, disability and work status only ethnic origin and age were identified by the classification tree algorithms as factors in separating successful from unsuccessful students. While the ethnicity was identified as one of the most influential factors among all predictors considered, age was important mostly in case of Level 5 courses. Those students under 30 taking level 5 courses were among those identified as the most vulnerable.

The only significant demographic factor was ethnic origin. Course related attributes such as course program and course block were also significant. However, these factors were not very accurate in identifying ‘at risk’ students. These results are consistent with other published research results. For example, Kotsiantis, Pierrakeas & Pintelas (2004) got similar prediction accuracy (between 58.84% when using neural network and 64.47% when using support vector machines) when only demographic variables were used. Background characteristics could be significant initially, i.e. taken as a group, but when other factors, related to the academic performance and environment, were included in the model, they dropped down on the rank list of important factors used for predicting study outcome.

5.5 Logistic regression

The logistic regression (binomial, or binary logistic regression) is a form of regression used when a dependent variable takes only two values (e.g. Study outcome 3 with two values: pass or fail). Multinomial (polytomous logistic regression) is used when a dependent variable has more classes than two (e.g. Study outcome 1 with three classes: *Pass*, *Fail* and *Lost*). Logistic regression could be used for the prediction of a study outcome and for determining the percentage of variation in the study outcome explained by the predictors (i.e. students demographics and course environment).

In the logistic regression analysis 37 variables, i.e. potential predictors were considered for each of three dependent variables of study outcome. Their definitions and reference categories are presented in Table 17. The fictional, reference student is male, under 30, disabled, Pacific Islander, with no secondary school qualification, not working, enrolled late, studying a level 5 course in online mode in semester 3 in the School of Business, and studying for a Bachelor of Applied Science. Table 13 and Table 40 present estimated binary and polytomous logistic regression models respectively showing estimated coefficients with their level of significance, odds ratios and a set of model diagnostics at the bottom. Odds ratio is used for interpretation of estimated logistic regression. Odds is the ratio of the probability something is true divided by the probability that it is not. Conditional odds is the ratio of probability something is true divided by the probability that it is not given the value of one of the variables. The odds ratio is the ratio of two odds or two conditional odds.

The *Odds ratio* column contains predicted changes in odds for a unit increase in the corresponding independent variable. Odds ratios less that 1 correspond to decreases in odds. Odds ratios greater than 1 correspond to increases in odds. Odds ratios close to 1 indicate that unit changes in that independent variable do not affect the dependent variable. In an attempt to measure the strength of association in a logistic regression various R^2 – like measures were proposed. Among them the Cox and Snell’s R^2 and Nagelkerke’s R^2 are the most reported. Because the Cox and

Snell’s R^2 can be less that 1.0 and difficult to interpret, Nagelkerke proposed further modification of the Cox and Snell’s R^2 to assure that it can vary from 0 to 1.

Table 13: Binary logistic regression model (Study outcome 3)

Independent variable	Study outcome 3	
	Coefficient	Odds ratio
Intercept	-3.761	
Student demographics		
Gender	0.251	1.285
Age group		
Between 30 and 40	0.507	1.660
Above 40	0.831	2.295
Disability	0.365	1.440
Ethnic group		
European	1.921	6.826
Chinese	1.521	4.577
Pakeha	1.588	4.894
Asian	1.225	3.404
Other	1.449	4.258
Indian	0.679	1.971
Māori	0.409	1.505
Secondary school		
NCEA Level 1	0.434	1.543
NCEA Level 2	0.733	2.080
University Entrance	0.851	2.343
NCEA Level 3	1.172	3.229
Overseas qualification	0.935	2.547
Other	0.502	1.653
Work status	0.231	1.260
Early enrolment	0.192	1.212
Course characteristics		
Course faculty		
School of Infor. and Social Sciences	0.582	1.790
Workplace Learning and Develop.	0.575	1.778
Course programme		
OP7001 Bachelor of Business	0.358	1.431
OP7020 Bachelor of Arts	0.572	1.772
Course level		
Level 6	0.929	2.532
Level 7	0.732	2.079
Course block		
Semester 1	0.301	1.352
Semester 2	0.167	1.182
Course offer type		
Distance	0.300	1.350
Blended	0.285	1.330
Number of observations	14704	
$-2 \log L$	14681.3	
Cox & Snell R^2	0.123	
Nagelkerke R^2	0.181	
Hosmer & Lemeshow test	8.510^{ns}	0.385
Overall % of correct classification	77.0%	

Note: Unless stated differently, all the coefficients are significant at less than the 1% level; *ns* stand for not significant.

The Hosmer-Lemeshow test of goodness of fit tests whether the model adequately fits the data. If the test is significant then the model does not adequately fit the data. However, the Hosmer-Lemeshow statistics does not accurately detect particular types of lack of fit, as noted by Agresti (2002).

From initial 37 potential predictors only 21 (Study outcome 1), 27 (study outcome 2) and 29 (study outcome 3) were identified as statistically significant. The most significant and also large in magnitude were the coefficients for the categories in the following predictors: ethnicity, course level, secondary school, age and course faculty. For example, holding other factors at a fixed value, the odds of a student being successful for European is 6 times over the odds of being successful for a Pacific Islands student (odds ratio is 6.826 for study outcome 3). In terms of percent change, we can say that the odds for European student are 583% higher than the odds for Pacific Islands student.

The odds of a student being successful if aged above 40 is 2.295 times over the odds of being successful for a student aged under 30. The odds of a student being successful for a student at work are 1.26 times over the odds of being successful for a student who is not working. As our results show, ethnicity, secondary school and the course level are on the top of the list of all predictors contributing the most to separation between successful and unsuccessful students in all three logistic regression models.

We are using the Nagelkerke’s R^2 coefficient and the Cox and Snell’s R^2 coefficient as a measure of association between study outcome and students demographics and course environment variables. They are taking the following values: 0.181 and 0.123 respectively in the logistic regression model for Study outcome 3. It means that only 18.1% of the variation in Study outcome 3 is explained by the independent variables. This also indicates that there are other factors not included in the logistic regression that explain the variation in the study outcome. It is believed that previous success is a good indicator of a future success. To check for the overall predictive accuracy of the logistic regression models reported in Table 13 and Table 40, classification matrix have been constructed for each of them. However, only the overall percentages are presented in the last row. The overall correct classification for Study outcome 3 was 77%. In other words the first model correctly predicts over 77% of the observations, classifying them correctly as a successful or unsuccessful student.

It is interesting to notice the differences in the odds ratios between *Pass* and *Lost* study outcomes in Table 40. Not all coefficients in the *Lost* column are significant and they are quite different from the corresponding odds ratios in the *Pass* column. That would suggest that different factors are contributing differently to these study outcomes and that the profile of those who decided to transfer or withdraw from the course is different from the profiles of students who passed or failed the course. This might be a topic for future research.

5.6 Discriminant analysis

Discriminant function analysis or discriminant analysis, is a multivariate statistical method used for the separation of groups. The goal of the discriminant analysis is to identify the relative contribution of variables to the separation between groups and to find an optimal separation between those groups. During the discriminant analysis, functions, (known as discriminant functions) are constructed based on the available variables in the data set that best describe separation between groups. Once the discriminant functions are constructed and the model evaluation shows that it is accurate in separating groups, we can use the discriminant functions to predict group membership for the data not used for constructing the model.

The discriminant analysis was carried out for all three study outcome variables. The results for study outcome 1 and 2 are presented in Table 41 and Table 43. Discriminant analysis for Study outcome 3 is presented in Table 15 with corresponding classification matrix presented in Table 14. Unstandardised canonical coefficients are used to make a decision on which group to classify the student into, the same way we have used regression coefficients in regression to make prediction. They are used for classification only. However, if we want to assess the relative importance of the independent variables we use standardised canonical coefficients.

Standardised canonical coefficients show the relative importance of the independent variables to enable the separation of successful and unsuccessful students. For example, from Table 15 for Study outcome 3 the variables that contribute the most to the separation of successful and unsuccessful students are: course level 5, course faculty – School of Business, course programme – Bachelor of Applied Science, secondary school qualification – NCEA Level 3, and age – under 30. The last two columns in these tables (column labelled: Wilks’ Lambda and *F*) are used to test which independent variable contribute significantly to the discriminant function. We read from Table 15, e.g. that Asian and Others ethnic groups are not contributing significantly to definition of the discriminant function.

Variables without standardised canonical coefficients didn’t pass the tolerance criteria and were not entered into the discriminant function. These are: age group (category: above 40), ethnicity (others), secondary school qualification (other), course level (level 7), course offer type (blended), course block (Semester 3), course programme (Bachelor of Arts) and course faculty (Workplace Learning and Development). The minimum tolerance limit was set to 0.001.

Structure coefficients (also called structure correlations) are the correlations between independent variables and the discriminant scores associated with a given discriminant function. For example from Table 15 (Study outcome 3) the highest correlation is between study outcome and the course level 5 (0.476). It is not a surprise that this independent variable has the highest standardised canonical coefficient and contributes the most to the separation of successful and unsuccessful students.

Table 14: Discriminant analysis classification matrix
(Study outcome 3)

	Predicted		
Observed	Fail	Pass	Percent correct
Fail	907	2801	24.5%
Pass	577	10419	94.8%
Overall percentage	61.1%	78.8%	77.0%

Table 15: Discriminant function summary (Study outcome 3)

Variable	Standardised canonical coefficient	Unstandardised canonical coefficient	Structure coefficient	Wilks Lambda	*F*
Intercept		-0.319			
Student demographics					
Gender	-0.125	-0.273	-0.111	0.998	27.23
Age group					
Under 30	0.422	0.981	0.359	0.981	286.48
Between 30 and 40	0.155	0.324	0.014	1.000	0.45[ns]
Above 40 [a]	-		-0.332	0.984	244.82
Disability	0.103	0.412	0.057	1.000	7.30
Ethnic group					
European	-0.086	-0.475	-0.098	0.999	21.42
Chinese	-0.016	-0.094	-0.032	1.000	2.24[ns]
Pakeha	-0.061	-0.140	-0.346	0.982	265.97
Asian	0.037	0.236	0.009	1.000	0.17[ns]
Other [a]	-		0.126	1.000	1.32[ns]
Indian	0.197	1.006	0.162	0.996	58.44
Māori	0.337	1.541	0.386	0.978	330.87
Pacific Islander	0.373	2.103	0.359	0.981	287.28
Secondary school					
No secondary school	0.177	0.679	0.339	0.983	255.69
NCEA Level 1	0.006	0.018	0.078	0.999	13.36
NCEA Level 2	-0.132	-0.333	0.020	1.000	0.92[ns]
University entrance	-0.191	-0.448	-0.139	0.997	43.16
NCEA Level 3	-0.283	-0.816	-0.131	0.997	38.12
Overseas qualification	-0.208	-0.537	-0.109	0.998	26.41
Other qualification [a]	-		-0.024	0.998	35.26
Work status	-0.119	-0.260	-0.168	0.996	62.89
Early enrolment	-0.086	-0.211	-0.145	0.997	46.93
Course characteristics					
Course faculty					
School of Business	0.347	0.697	0.086	0.999	16.56
School of ISS	0.009	0.018	-0.089	0.999	17.78
Workplace Learning [a]	-		0.009	1.000	0.17[ns]
Course programme					
Bachelor of Business	0.066	0.134	0.055	1.000	6.74
Bachelor of Appl. Sci.	0.265	0.577	0.059	0.999	7.83
Bachelor of Arts [a]	-		-0.167	0.996	62.00
Course level					
Level 5	0.416	0.845	0.476	0.967	503.39
Level 6	-0.081	-0.180	-0.339	0.983	255.92
Level 7 [a]	-		-0.201	0.994	90.12
Course block					
Semester 1	-0.185	-0.373	-0.147	0.997	47.80
Semester 2	-0.111	-0.223	0.000	1.000	0.01[ns]
Semester 3 [a]	-		0.219	0.993	106.78
Course offer type					
Distance	-0.004	-0.009	-0.179	0.995	70.94
Online	0.129	0.388	0.174	0.995	67.53
Blended [a]	-		0.055	1.000	6.66

Note: Unless stated differently, all the coefficients are significant at less than the 1% level; *ns* stand for not significant.

[a] This variable is not used in the analysis

The conclusion based on the results in Table 15 for Study outcome 3 is similar to conclusion that could be made based on discriminant analysis of Study outcome 1 and 2. The list of the most important factors is the same with quite similar relative contributions to the separation between the groups. Therefore we can conclude that the results of the discriminant analysis are quite robust to changes in the definition of the study outcome.

Table 16 summarises the key features of the discriminant functions. The eigenvalue shows the relative importance of the discriminant function if there is more than one function. The canonical correlation of each discriminant function is the correlation of that function with the discriminant scores, values resulting from applying a discriminant function formula to the data.

The coefficient of determination, i.e. the squared canonical correlation, is a percentage of the variation in the study outcome discriminated by the set of independent variables. The highest percentage of variations is obtained with the discriminant function for Study outcome 3 (13.1%). In other words only 13.1% variation in this study outcome is discriminated by the predictors.

The Wilks’ Lambda is used to test the significance of the discriminant function as a whole. The chi-square statistics with given degree of freedom (*d.f.*) in Table 16 show that all discriminant functions are statistically significant at the 1% level.

Table 16: Discriminant functions summary

Dependent variable	Eigenvalue	Canonical correlation	Coefficient of determination	Wilks’ Lambda	χ^2	*d.f.*
Study outcome 1						
First function	0.117	0.323	0.1043	0.886	2349.39	58
Second function	0.010	0.102	0.0104	0.990	202.50	28
Study outcome 2						
First function	0.095	0.295	0.0870	0.913	1768.31	29
Study outcome 3						
First function	0.151	0.362	0.1310	0.869	2068.85	29

The problem with the use of discriminant analysis in retention studies is that the assumptions the model is based on, are not always satisfied. Therefore we tested a hypothesis about the homogeneity of covariances, i.e. covariance matrices do not differ between groups (e.g. *Pass* and *Fail*). The Box’s M test was used to test the hypothesis that the covariance matrices are equal. This hypothesis was rejected (test results are not presented). However, in large samples even small differences in covariance matrices may be found significant by Box’s M, when in fact there is no problem of violation of this assumption. The significant Box’s M could be ignored when the group log determinants are similar. This was the case for all three study outcome variables. The discriminant analysis also assumes that we are dealing with interval data. However, all the independent variables in our case are dummy, i.e. binary variables taking values 0 and 1 only. Though the studies show that the violation of these assumptions is not quite crucial for the classification accuracy of the discriminant functions, there is always a chance that the results will be biased due to a departure from these assumptions.

6. Concluding Remarks

This study examines the background information from enrolment data that impacts upon the study outcome of students at the Open Polytechnic.

Generally speaking more students are lost by attrition than failure i.e. they enrol and either withdraw, are academically withdrawn or transfer to the next semester. This suggests that engagement is a problem for students with little external motivation. This may be the result of years of face-to-face education where external motivation by a teacher is high, unlike the business world where internal motivation is a necessary requirement for advancement and success. A pre-test of internal motivators may be a key hurdle for students to surmount. Just taking a pre-test that requires significant thought and input may be a sufficient indicator of the student’s internal motivation.

Indian, Māori and Pacific students have lower pass rates than the general population with comparable *Lost* rates. These students are therefore retained but fail to reach the required standard with a pass rate half that of the rest of the population. This is particularly seen amongst those under 40 on level 5 courses, a trend also seen in the rest of the population. The pass rate rises amongst Māori on courses at level 6 and 7 but is still substantially lower than the rest of the population.

Of those in the general population studying at level 6 and 7 there is a greater chance of failing or being lost amongst those who are not working in spite of a much greater pass rate amongst those studying at this level.

Students under 30 from the general population studying at level 5 are more likely to pass a Bachelor of Arts course than a Bachelor of Applied Science or Business degree course, though more students study the latter two courses.

Of the general population over 40 studying at level 5, previous study at any level is a significant indicator of future success. Those with no previous study are far more likely to be lost and twice as likely to fail.

Based on our results the most important factors that help separate successful from unsuccessful students are ethnicity, course level, secondary school, age, course programme and course block. More specifically, the most vulnerable students are Pacific Islands, Māori and marginally Indian students, those studying level 5 courses, with no secondary school qualification, being under 30, enrolled in a Bachelor of Applied Science programme and studying in Semester 3. Other factors, such as gender, work status and early enrolment also appear in some of the models but they are ranked lower on the importance list. These results are consistent with the results obtained in the previous studies. For diploma level courses at the Open Polytechnic Bathurst (2004) also identified Pacific Islands and Māori as ‘at risk’ students particularly those with minimal or no secondary school qualifications. In the similar study for the Open University in UK Woodman (2001) listed ethnicity, course level, age and previous education among significant factors for study outcome. Simpson (2006) found that the course level, previous education and course programme are important factors determining study outcome of the newly registered students at the Open University in UK. Finally, Herrera (2006) identified a programme level as one of the significant factors for predicting student persistence.

The classification accuracy varies between models. The logistic regression and discriminant analysis models achieved higher overall classification accuracy than the classification tree models (between 1% to 4%), but at the cost of using in some cases up to 8 times more variables. The CART classification trees were slightly more accurate than the CHAID trees and were also more parsimonious models than the CHAID trees and even more than the logistic regression and discriminant analysis models. If two models explain equally well some phenomenon, then Occam's razor recommends the selection of the model that uses fewer variables, or has fewer parameters. Therefore we would recommend the use of the CART classification tree model in the early identification of 'at risk' students.

A drop in the overall accuracy of the four models was noticed for the study outcome variables (Study outcome 1 and Study outcome 2) that include students who transferred or withdrew (academically or voluntarily). In other words, greater accuracy of the models was achieved excluding *Lost* students (i.e. transferees and withdrawals). For example in the case of the discriminant analysis model, accuracy from 77% in the case of Study outcome 3 variable dropped to 64% in the case of Study outcome 2 variable and further to 58.3% in the case of Study outcome 1 variable. Study outcome 2 variable merged them into one group together with those students who failed the course, while Study outcome 1 variable separates students into three groups labelled as *Pass*, *Fail* and *Lost*. These results confirm that the transferees and withdrawals should be modelled as a separate group. For future research it would be interesting to investigate further how the students who transferred or withdrew (academically or voluntarily) are different from those who failed the course. The conclusions and achieved accuracy level in our study are comparable with the accuracy levels obtained and conclusion reached in the previous studies. Kember (1995) found that generally background information is not a good predictor of the final, study outcome. Kotsiantis, Pierrakeas & Pintelas (2004) found that when only the demographic variables were used, accuracy level was under 65%. Even lower accuracy level was obtained in the Vandamme, Meskens & Superby (2007) study: only 40% when using decision, i.e. classification trees and about 57% with the discriminant analysis.

The overall classification accuracy was reasonably high in the case of two groups (excluding transferees and withdrawals from models) and at the same level achieved in other research studies where only the enrolment data was used. The lower level of overall classification accuracy in the case of Study outcome 2 variable suggests that unsuccessful students (*Fail*) and transferees and withdrawals (*Lost*) should not be modelled as one homogeneous population. They seem to have different reasons for not completing the course in the first attempt and therefore should be modelled as a separate group. However, when we modelled them as a separate group (Study outcome 3 variable) the overall classification accuracy dropped further. This would suggest that the student demographics (gender, age, ethnicity, disability, secondary school, work status, and early enrolment) and course characteristics (faculty, programme, level, block and offer type) gathered during the enrolment process do not contain sufficient information for an accurate separation of successful, unsuccessful and transferees and withdrawals students.

Our results have some interesting practical implications for both academic and administrative staff at the Open Polytechnic. Classifying students based on pre-enrolment information and the rules presented for each node in the classification tree would allow the administrative and academic staff to identify students who would be 'at risk' of dropping the course even before they start with their study. Then the student support systems, such as orientation, advising, and mentoring programs or tertiary study skills courses or compulsory pre-tests, could be used to positively impact the academic successes of such students.

This study is limited in three main ways that future research can perhaps address. Firstly, our research is based on enrolment data only. Leaving out other important factors (academic achievement, number of courses completed, motivation, financial aids, etc.) that may affect study outcome and could distort results obtained with models used. For example, including the assignment mark after the submission of the first course assignment or even better a pre-entry test would probably improve the predictive accuracy of the models. To improve the model, more attributes could be included to obtain prediction models with lower misclassification errors. However, the model in this case would not be a tool for pre-enrolment, i.e. early identification of 'at risk' students.

Secondly, the time line should be included in the analysis. We would need to follow those students who failed the course and also transferees and withdrawals. Some of them may re-enrol in one of the next semesters and might successfully complete the course in the second or third attempt. Tracking *Fail* and *Lost* students in subsequent semesters and tracking their study outcome would help modelling their behaviour more accurate.

Thirdly, from a methodological point of view an alternative to logistic regression and discriminant analysis should be considered. The prime candidate to be used with this data set is neural networks. We may also consider other classification tree models such as exhaustive CHAID, QUEST, random forest, and ensembles of models.

Finally, besides enhancing the accuracy of prediction one of the directions for future research could be focused on using the data currently collected to identify the best support systems.

7. Acknowledgements

This research project was financially supported by Ako Aotearoa. We wish also to thank Mr Cheng Qian for extracting data from Integrator, the student management system of the Open Polytechnic.

8. References

Agresti, A. (2002). *Categorial data analysis* (2nd ed.), New York: John Wiley and Sons.

Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006). Mining student data using decision trees. In the *Proceedings of the 2006 International Arab Conference on Information Technology (ACIT'2006)*.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*, 3-17.

Bathurst, J. (2004). An analysis of Diploma of Health and Human Behaviour completions 2002. *Working Paper No. 2-04*. Wellington: Open Polytechnic.

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55*, 485-540.

Boero, G., Laureti, T., & Naylor, R. (2005). An econometric analysis of student withdrawal and progression in post-reform Italian universities. Centro Ricerche Economiche Nord Sud - *CRENoS Working Paper 2005/04*.

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In the *Proceedings of 5th Annual Future Business Technology Conference*, Porto, Portugal, 5-12.

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting student drop out: A case study. In the *Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09)*. July 1-3, Cordoba, Spain, 41-50.

Dirkx, J. M., & Jha, L. R. (1994). Completion and attrition in adult basic education: A test of two pragmatic prediction models. *Adult Education Quarterly*, *45*(1), 269-285.

Dupin-Bryant, P. A. (2004). Pre-entry variables related to retention in online distance education. *The American Journal of Distance Education, 18*(4), 199-206.

Glynn, J. G., Sauer, P. L., & Miller, T. E. (2003). Signaling student retention with prematriculation data. *NASPA Journal, 41*(1), 41- 67.

Grote, B. (2000). Student retention and support in open and distance learning. *Working Paper No. 2-00*. Wellington: Open Polytechnic.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.), Amsterdam: Elsevier.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.), New York: Springer.

Herrera, O. L. (2006). *Investigation of the role of pre- and post-admission variables in undergraduate institutional persistence, using a Markov student flow model*. PhD Dissertation, North Carolina State University, USA.

Horstmanshof, L., & Zimitat, C. (2007). Future time orientation predicts academic engagement among first-year university students. *British Journal of Educational Psychology, 77* (3): 703-718.

Ishitani, T. T. (2003). A longitudinal approach to assessing attrition behavior among first-generation students: Time-varying effects of pre-college characteristics. *Research in Higher Education, 44*(4), 433-449.

Ishitani, T. T. (2006). Studying attrition and degree completion behavior among first-generation college students in the United States. *Journal of Higher Education, 77*(5), 861-885.

Jun, J. (2005). *Understanding dropout of adult learners in e-learning*. PhD Dissertation, The University of Georgia, USA.

Kember, D. (1995). *Open learning courses for adults: A model of student progress*. Englewood Cliffs, NJ: Education Technology.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence, 18,* 411-426.

Luan, J., & Zhao, C-M. (2006). Practicing data mining for enrollment management and beyond. *New Directions for Institutional Research, 31*(1), 117-122.

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

Nandeshwar, A., & Chaudhari, S. (2009). Enrollment prediction models using data mining. Retrieved January 10, 2010, from http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Amsterdam: Elsevier.

Noble, K., Flynn, N. T., Lee, J. D., & Hilton, D. (2007). Predicting successful college experiences: Evidence from a first year retention program. *Journal of College Student Retention: Research, Theory & Practice,* 9(1), 39-60.

Pascarella, E. T., Duby, P. B., & Iverson, B. K. (1983). A test and reconceptualization of a theoretical model of college withdrawal in a commuter institution setting. *Sociology of Education, 56*, 88-100.

Pratt, P. A., & Skaggs, C. T. (1989). First-generation college students: Are they at greater risk for attrition than their peers. *Research in Rural Education, 6*(1), 31-34.

Rokach, L., & Maimon, O. (2008). *Data mining with decision trees – Theory and applications*. New Jersey: World Scientific Publishing.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33,* 135-146.

Simpson, O. (2006). Predicting student success in open and distance learning. *Open Learning, 21*(2), 125-138.

Siraj, F., & Abdoulha, M. A. (2009). Uncovering hidden information within university's student enrolment data using data mining. *MASAUM Journal of Computing, 1*(2), 337-342.

Strayhorn, T. L. (2009). An examination of the impact of first-year seminars on correlates of college student retention. *Journal of The First-Year Experience & Students in Transition, 21*(1), 9-27.

Tharp, J. (1998). Predicting persistence of urban commuter campus students utilizing student background characteristics from enrollment data. *Community College Journal of Research and Practice, 22*, 279-294.

Vandamme, J.-P., Meskens, N., & Superby, J.-F. (2007). Predicting academic performance by data mining methods. *Education Economics, 15*(4), 405-419.

Woodman, R. (2001). *Investigation of factors that influence student retention and success rate on Open University courses in the East Anglia region*. M.Sc. Dissertation, Sheffield Hallam University, UK.

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., Lo, W., & Kaprolet, C. (2007). A data-mining approach to differentiate predictors of retention. In the *Proceedings of the Educause Southwest Conference,* Austin, Texas, USA.

9. Appendix A: Data description

Table 17: Description of variables and their domains

Variable	Description (Domain; Reference group in *italic*)
	Student demographics
Gender	Student gender (binary: female or *male*)
Age	Student’s age (numeric: 1 – *under 30*, 2 – 30 to 40 or 3 – over 40)
Age 1	Student’s age under 30 (binary: yes or no)
Age 2	Student’s age between 30 to 40 (binary: yes or no)
Age 3	Student’s age above 40 (binary: yes or no)
Pakeha	Student belongs to NZ European / Pakeha ethnic group (binary: yes or no)
Māori	Student belongs to NZ Māori ethnic group (binary: yes or no)
Pacific	Student belongs to Pacific Islands Māori ethnic group that includes: Samoan, Cook Island Māori, Tongan, Niuean, Tokelauan, Fijian and other Pacific peoples (binary: yes or no)
Indian	Student belongs to Indian ethnic group (binary: yes or no)
Chinese	Student belongs to Chinese ethnic group (binary: yes or no)
European	Student belongs to ‘European’ ethnic group that includes: British/Irish, Dutch, Greek, Polish, South Slav, Italian, German and other European (binary: yes or no)
Asian	Student belongs to ‘Asian’ ethnic group that includes: Filipino, Cambodian, Vietnamese, other South Asian, Sri Lankan, Japanese, Korean and other Asian (binary: yes or no)
Others	Student is classified as neither Pakeha, Māori, Pacific, Indian, Chinese, European or Asian (binary: yes or no)
Ethnicity	Student’s ethnic group (nominal: Pakeha, Māori, *Pacific*, Indian, Chinese, European, Asian or Others)
Disability	Student has a disability (binary: yes or *no*)
	Pre-enrolment experience
Work status	Student is working (binary: *yes* or no)
Secondary school	Student’s highest level of achievement from a secondary school (nominal: *No secondary qualification*, NCEA1, NCEA2, University entrance, NCEA3, Overseas or Others)
No secondary qualification	Student has no formal secondary school qualification (binary: yes or no)
NCEA1	Student achieved NCEA Level 1 or School Certificate (binary: yes or no)
NCEA2	Student achieved NCEA Level 2 or 6th Form Certificate (binary: yes or no)
University entrance	Student achieved University Entrance (binary: yes or no)
NCEA3	Student achieved NCEA Level 3 or Bursary or Scholarship (binary: yes or no)
Overseas	Student achieved Overseas qualification (binary: yes or no)
Other	Student achieved Other qualification (binary: yes or no)

Table 17: Description of variables and their domains

Variable	Description (Domain; Reference group in *italic*)
Early enrolment	Student enrolled for the first time in the course before start of the course (binary: yes or *no*)
	Study environment
Course level	Course level (nominal: *5*, 6 or 7)
Course level 5	Course Level 5 (binary: yes or no)
Course level 6	Course Level 6 (binary: yes or no)
Course level 7	Course Level 7 (binary: yes or no)
Course faculty	Course Faculty (nominal: *School of Business*, School of Information and Social Sciences and Workplace Learning and Development)
Course faculty 1	Course Faculty School of Business (binary: yes or no)
Course faculty 2	Course Faculty School of Information and Social Sciences (binary: yes or no)
Course faculty 3	Course Faculty Workplace Learning and Development (binary: yes or no)
Course programme	Programme (nominal: OP7001 – Bachelor of Business, *OP7010 – Bachelor of Applied Science* or OP7020 – Bachelor of Arts)
Course programme 1	Programme: OP7001 – Bachelor of Business (binary: yes or no)
Course programme 2	Programme: OP7010 – Bachelor of Applied Science (binary: yes or no)
Course programme 3	Programme: OP7020 – Bachelor of Arts (binary: yes or no)
Course offer type	Course offer type (nominal: *Online*, Distance or Blended)
Course offer type 1	Course offer type: Online (binary: yes or no)
Course offer type 2	Course offer type: Distance (binary: yes or no)
Course offer type 3	Course offer type: Blended (binary: yes or no)
Course block	Semester in which a course is offered (Semester 1, Semester 2 or *Semester 3*)
Course block 1	Course is offered in Semester 1 (binary: yes or no)
Course block 2	Course is offered in Semester 2 (binary: yes or no)
Course block 3	Course is offered in Semester 3 (binary: yes or no)
	Dependent variable
Study outcome 1 Pass/Fail/Lost	Study outcome (nominal: Pass – successful completion, *Fail – unsuccessful completion* and Lost – withdrawal, academic withdrawal and transfer)
Study outcome 2 Pass/Fail (includes Lost)	Study outcome (binary: Pass – successful completion, *Fail – unsuccessful completion* includes also Lost – withdrawal, academic withdrawal and transfer)
Study outcome 3 Pass/Fail (excludes Lost)	Study outcome (binary: Pass – successful completion, *Fail – unsuccessful completion* excludes Lost – withdrawal, academic withdrawal and transfer)

10. Appendix B: Glossary

At risk students Students whose characteristics (biological, socio-economical, and other factors) might increase a probability of not completing the course / diploma / degree.

Attrition The number of students not completing their current semester of enrolment. Students who finish semester with a grade A, B, C, D, or F are considered to have completed, while students who are flagged W (withdraw), T (transfer) or AW (academic withdraw) are considered not to have completed the coursework.

Dropouts Students who discontinue their enrolment and do not re-enrol with the Open Polytechnic to continue their study.

First timers Students who enrolled with the Open Polytechnic for the first time.

Late comers Students who enrolled with the Open Polytechnic in the week before enrolments close or whose enrolment form was processed after the course started. Due to the enrolment process they will get course material or access to the course web page and a set text after the course started.

Persistence The willingness (conscious decision) of students to continue their study and successfully complete the course / diploma / degree.

Resilience Students who persist despite having ‘at risk’ conditions (socio-economic, demographic and other factors)

Retention Students returning to the Open Polytechnic after their first semester/course of enrolment, as well as for subsequent semesters/courses. Retention occurs when the Open Polytechnic successfully supports student persistence.

11. Appendix C: Charts for variables used

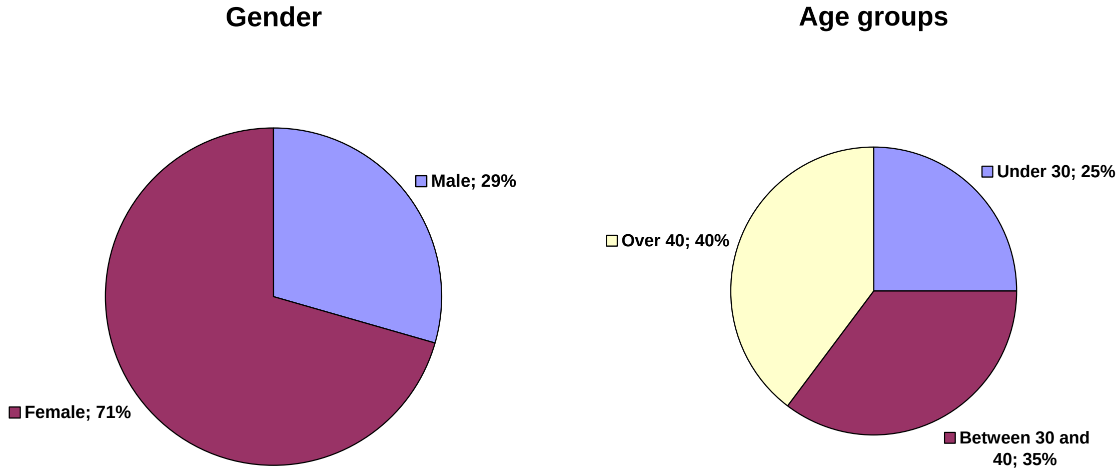


Figure 8: Students by Gender and Age

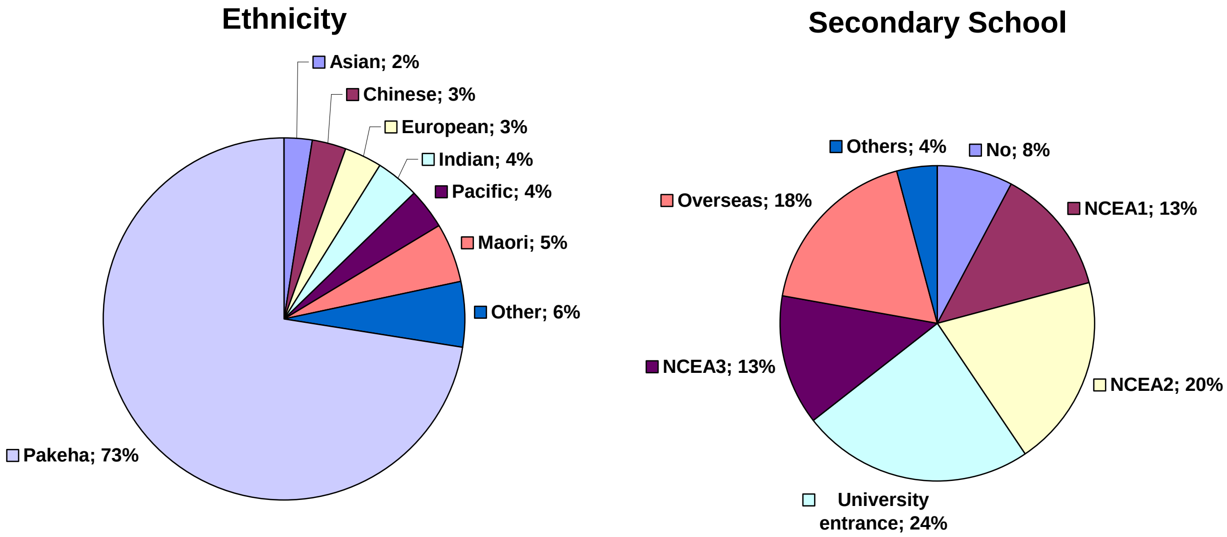


Figure 9: Students by Ethnicity and Secondary school

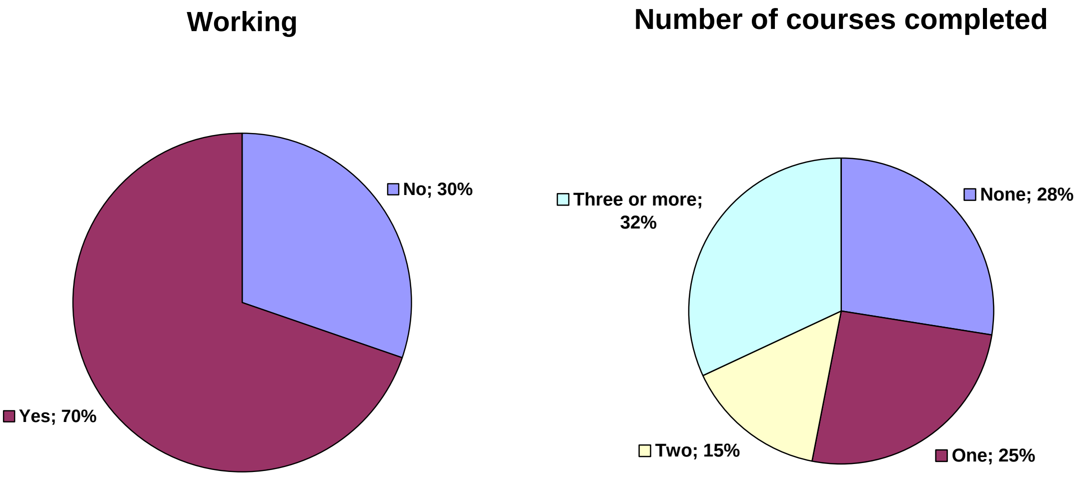


Figure 10: Students by the Work status and Number of courses completed

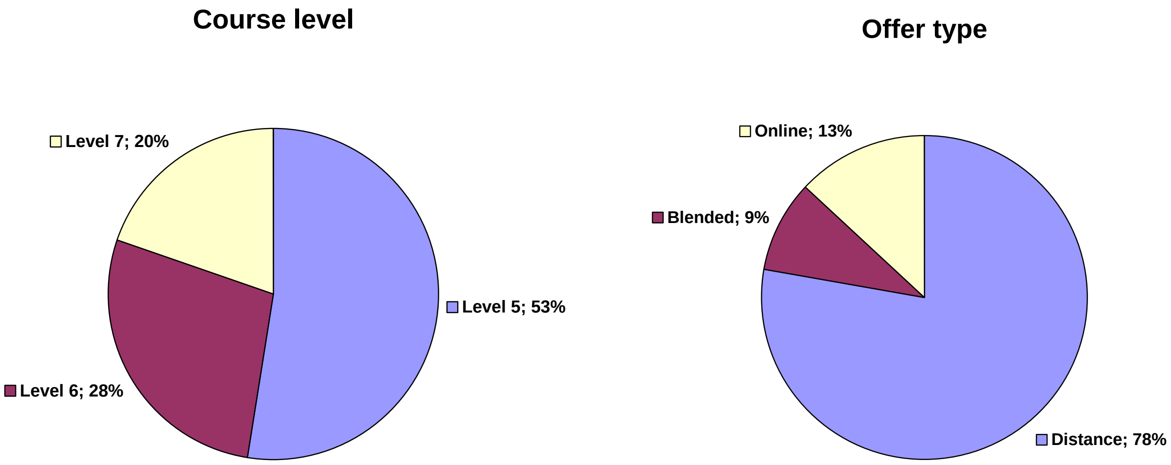


Figure 11: Students by the Course level and Course offer type

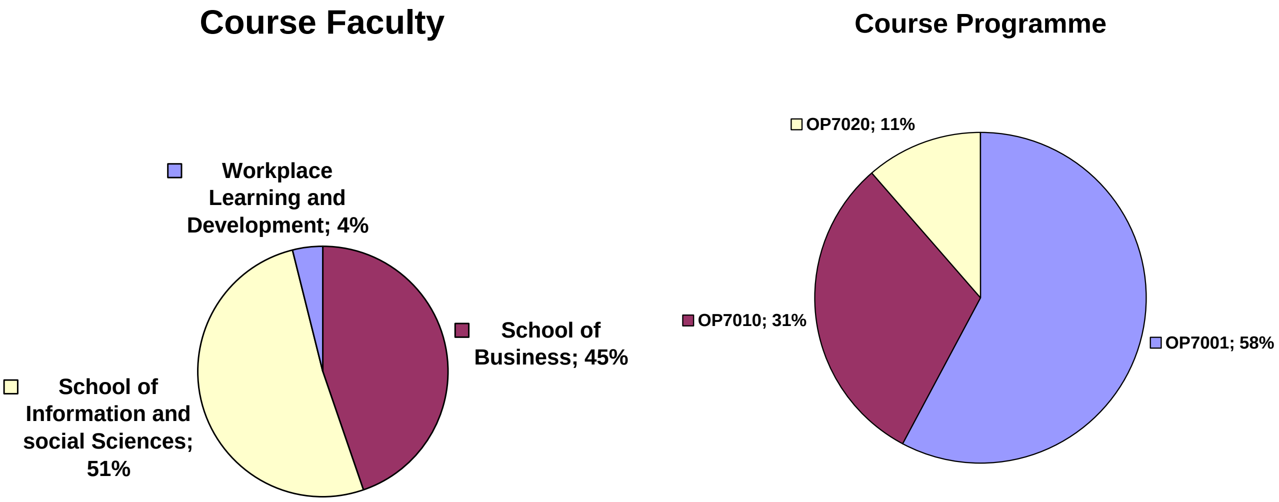


Figure 12: Students by the Course faculty and Course programme

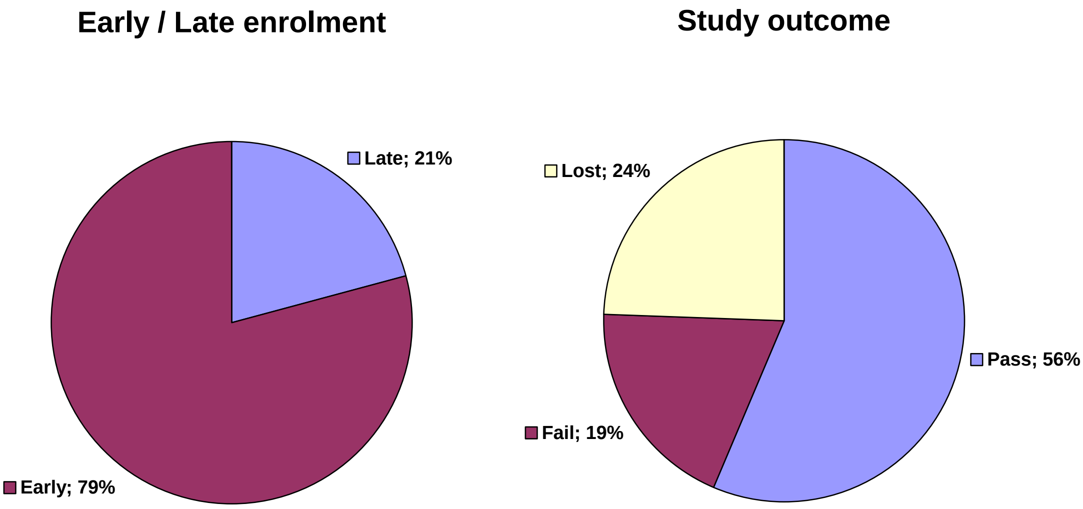


Figure 13: Students by Early enrolment and Study outcome

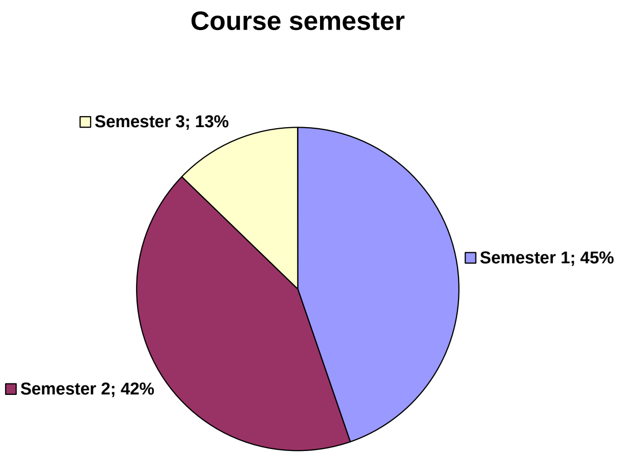
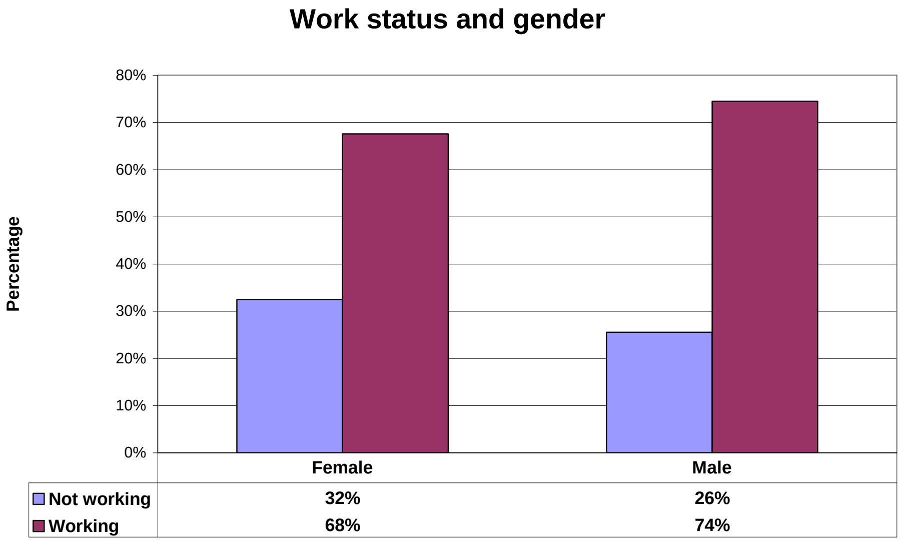
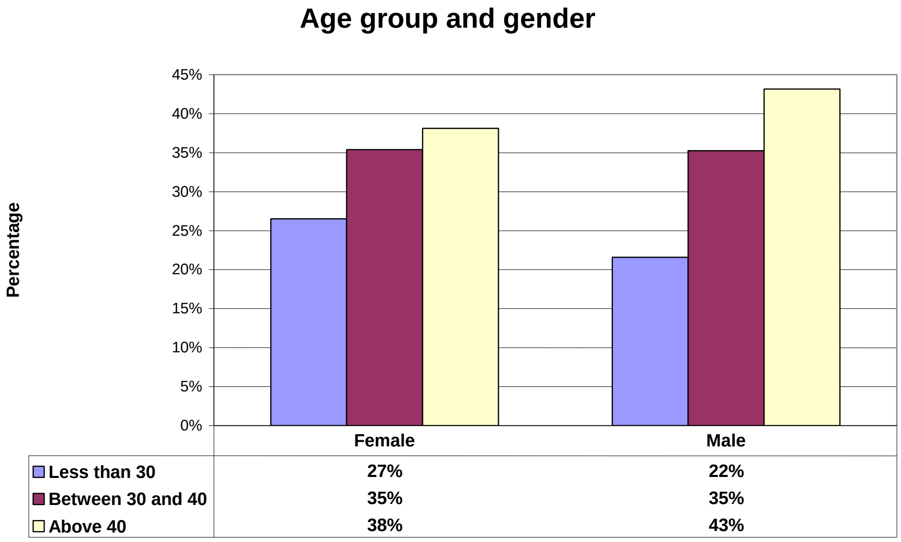


Figure 14: Students by the Course block



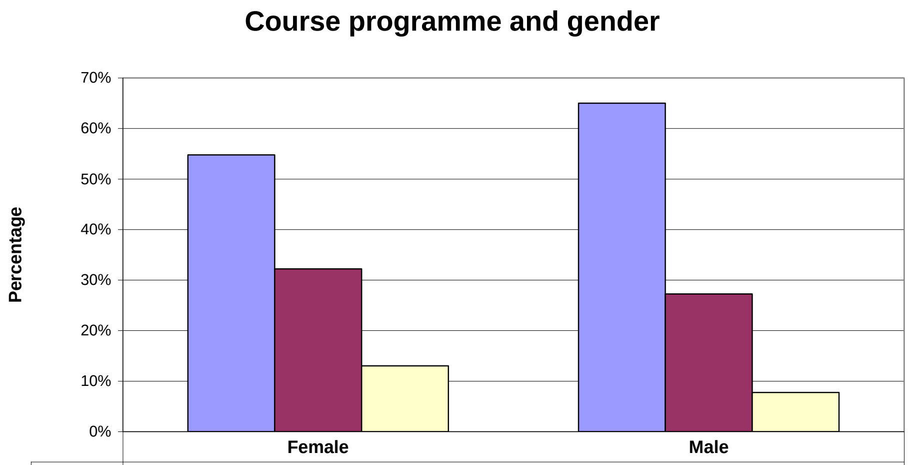
	Female	Male
□ Not working	32%	26%
■ Working	68%	74%

Figure 15: Study Outcome by Work status and Gender



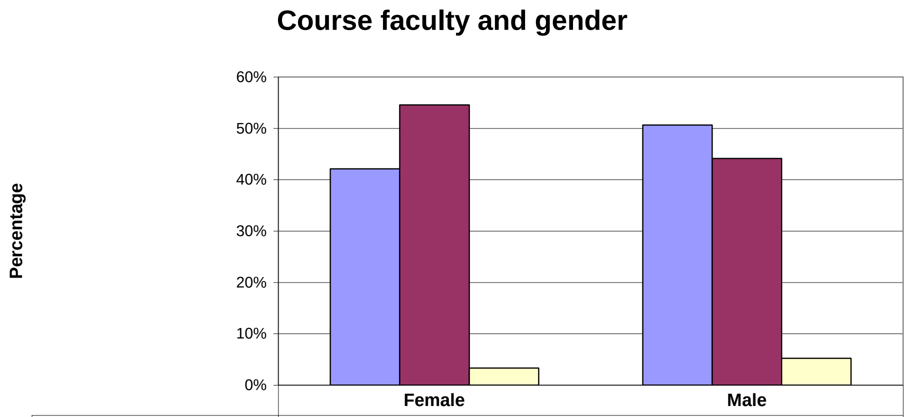
	Female	Male
□ Less than 30	27%	22%
■ Between 30 and 40	35%	35%
□ Above 40	38%	43%

Figure 16: Study Outcome by Age and Gender



	Female	Male
■ OP7001	55%	65%
■ OP7010	32%	27%
□ OP7020	13%	8%

Figure 17: Study Outcome by Course programme and Gender



	Female	Male
■ School of Business	42%	51%
■ School of Information and Social Sciences	55%	44%
□ Workplace Learning and Development	3%	5%

Figure 18: Students by Course faculty and Gender

12. Appendix D: Classification trees and rules

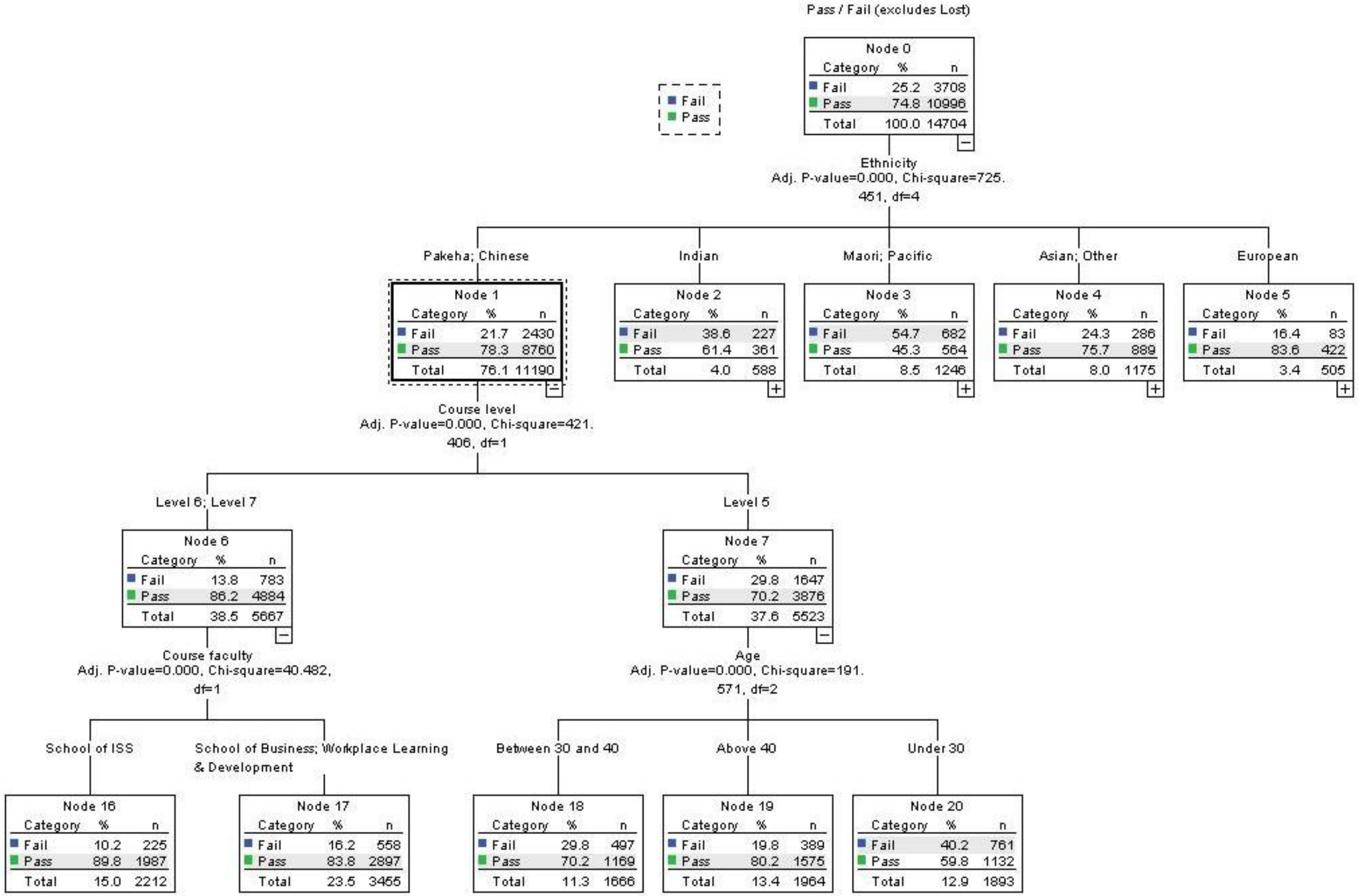


Figure 19: CHAID tree (Study outcome 3: Pakeha & Chinese)

Table 18: Rules for CHAID tree (Study outcome 3: Pakeha & Chinese)

Node	Rule	Outcome	Probability
16	**IF** Ethnicity = “Pakeha” **OR** “Chinese” **AND** Course level = “Level 6” **OR** “Level 7” **AND** Course faculty = “School of Information and Social Sciences” **THEN**	Pass	0.898
17	**IF** Ethnicity = “Pakeha” **OR** “Chinese” **AND** Course level = “Level 6” **OR** “Level 7” **AND** Course faculty = “School of Business **OR** “Workplace Learning and Development” **THEN**	Pass	0.838
18	**IF** Ethnicity = “Pakeha” **OR** “Chinese” **AND** Course level = “Level 5” **AND** Age = “Between 30 and 40” **THEN**	Pass	0.702
19	**IF** Ethnicity = “Pakeha” **OR** “Chinese” **AND** Course level = “Level 5” **AND** Age = “Above 40” **THEN**	Pass	0.802
20	**IF** Ethnicity = “Pakeha” **OR** “Chinese” **AND** Course level = “Level 5” **AND** Age = “Under 30” **THEN**	Pass	0.598

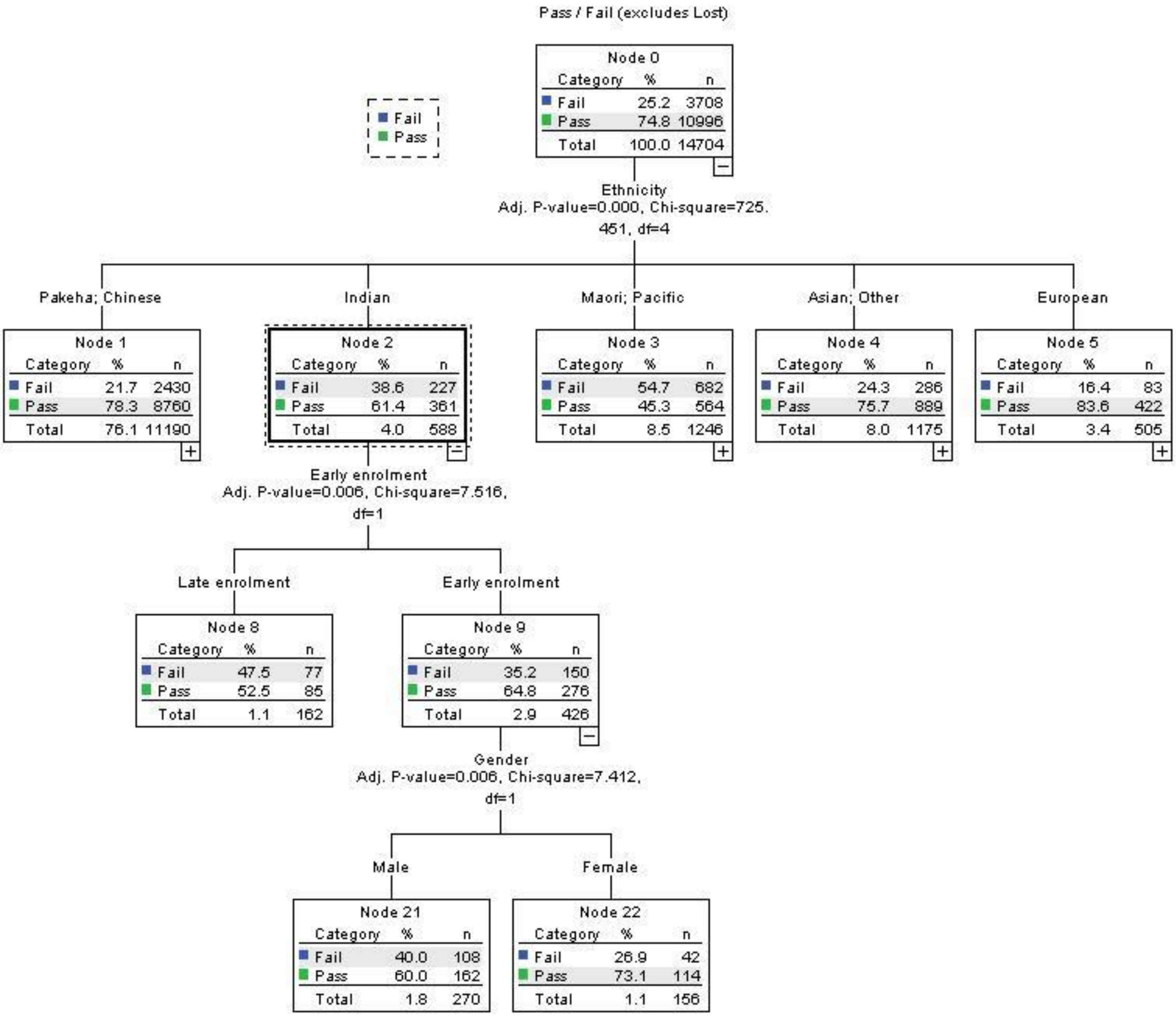


Figure 20: CHAID tree (Study outcome 3: Indian)

Table 19: Rules for CHAID tree (Study outcome 3: Indian)

Node	Rule	Outcome	Probability
8	**IF** Ethnicity = "Indian" **AND** Early enrolment = "Late" **THEN**	Pass	0.525
21	**IF** Ethnicity = "Indian" **AND** Early enrolment = "Early" **AND** Gender = "Male" **THEN**	Pass	0.600
22	**IF** Ethnicity = "Indian" **AND** Early enrolment = "Early" **AND** Gender = "Female" **THEN**	Pass	0.731

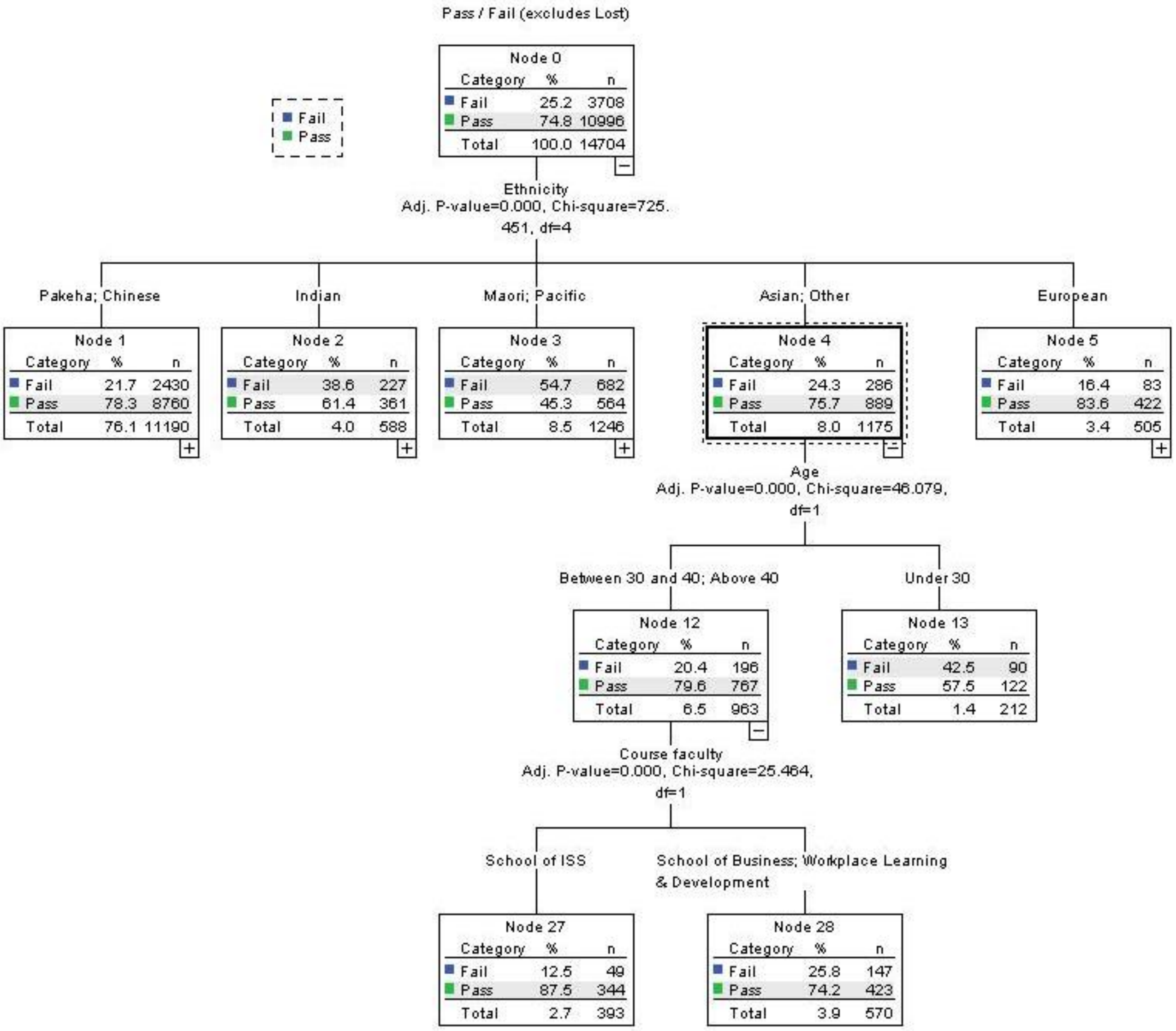


Figure 21: CHAID tree (Study outcome 3: Asian & Others)

Table 20: Rules for CHAID tree (Study outcome 3: Asian & Others)

Node	Rule	Outcome	Probability
13	**IF** Ethnicity = "Asian" **OR** "Other" **AND** Age = "Under 30" **THEN**	Pass	0.575
27	**IF** Ethnicity = "Asian" **OR** "Other" **AND** Age = "Between 30 and 40" **OR** "Above 40" **AND** Course faculty = "School of Information and Social Sciences" **THEN**	Pass	0.875
28	**IF** Ethnicity = "Asian" **OR** "Other" **AND** Age = "Between 30 and 40" **OR** "Above 40" **AND** Course faculty = "School of Business" **OR** "Workplace Learning and Development" **THEN**	Pass	0.742

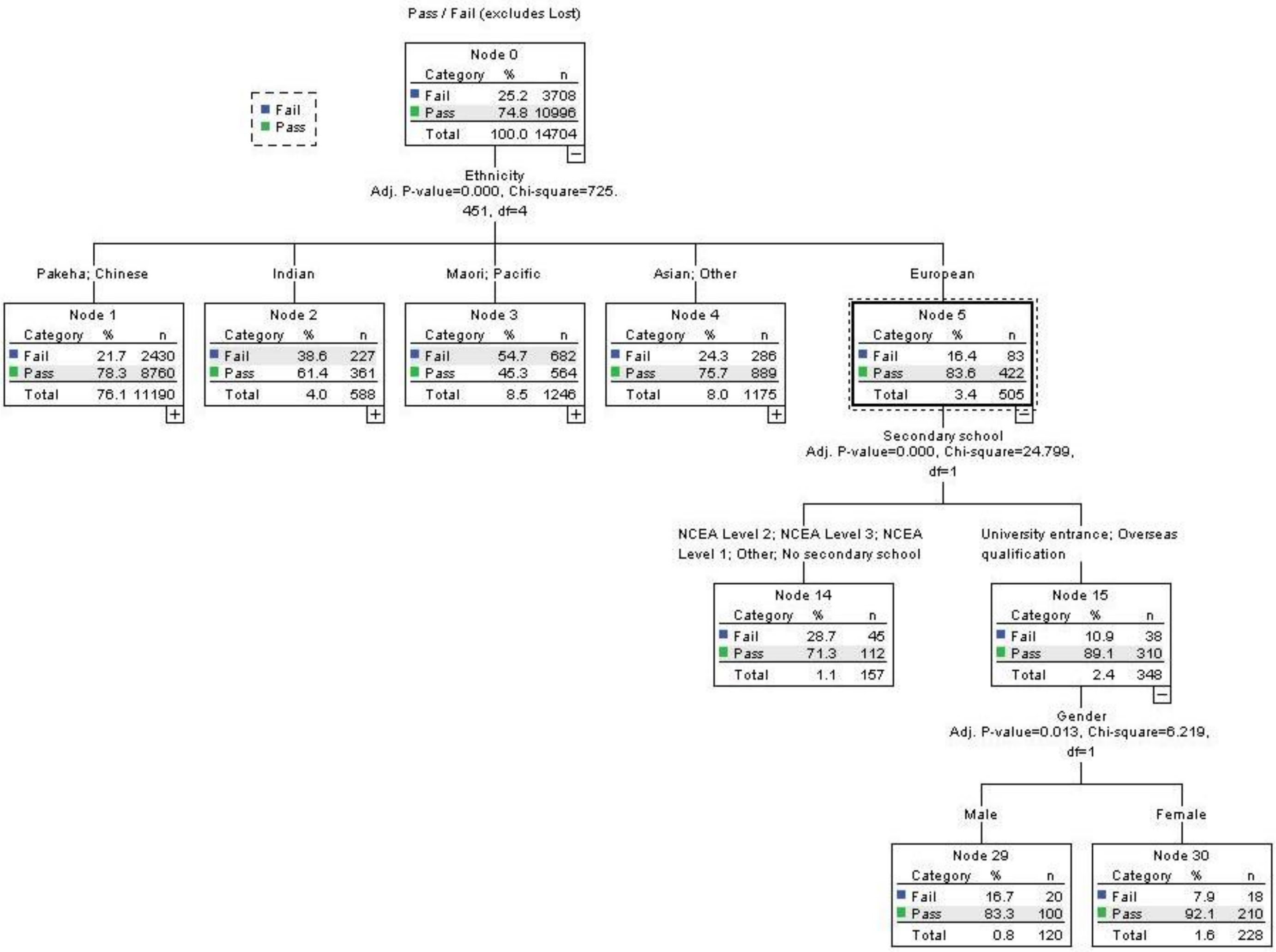


Figure 22: CHAID tree (Study outcome 3: European)

Table 21: Rules for CHAID tree (Study outcome 3: European)

Node	Rule	Outcome	Probability
14	**IF** Ethnicity = “European” **AND** Secondary school = “No secondary school” **OR** “NCEA Level 1”**OR** “NCEA Level 2”**OR** “NCEA Level 3”**OR** “Other” **THEN**	Pass	0.575
29	**IF** Ethnicity = “European” **AND** Secondary school = “University entrance” **OR** “Overseas qualification” **AND** Gender = “Male” **THEN**	Pass	0.833
30	**IF** Ethnicity = “European” **AND** Secondary school = “University entrance” **OR** “Overseas qualification” **AND** Gender = “Female” **THEN**	Pass	0.921

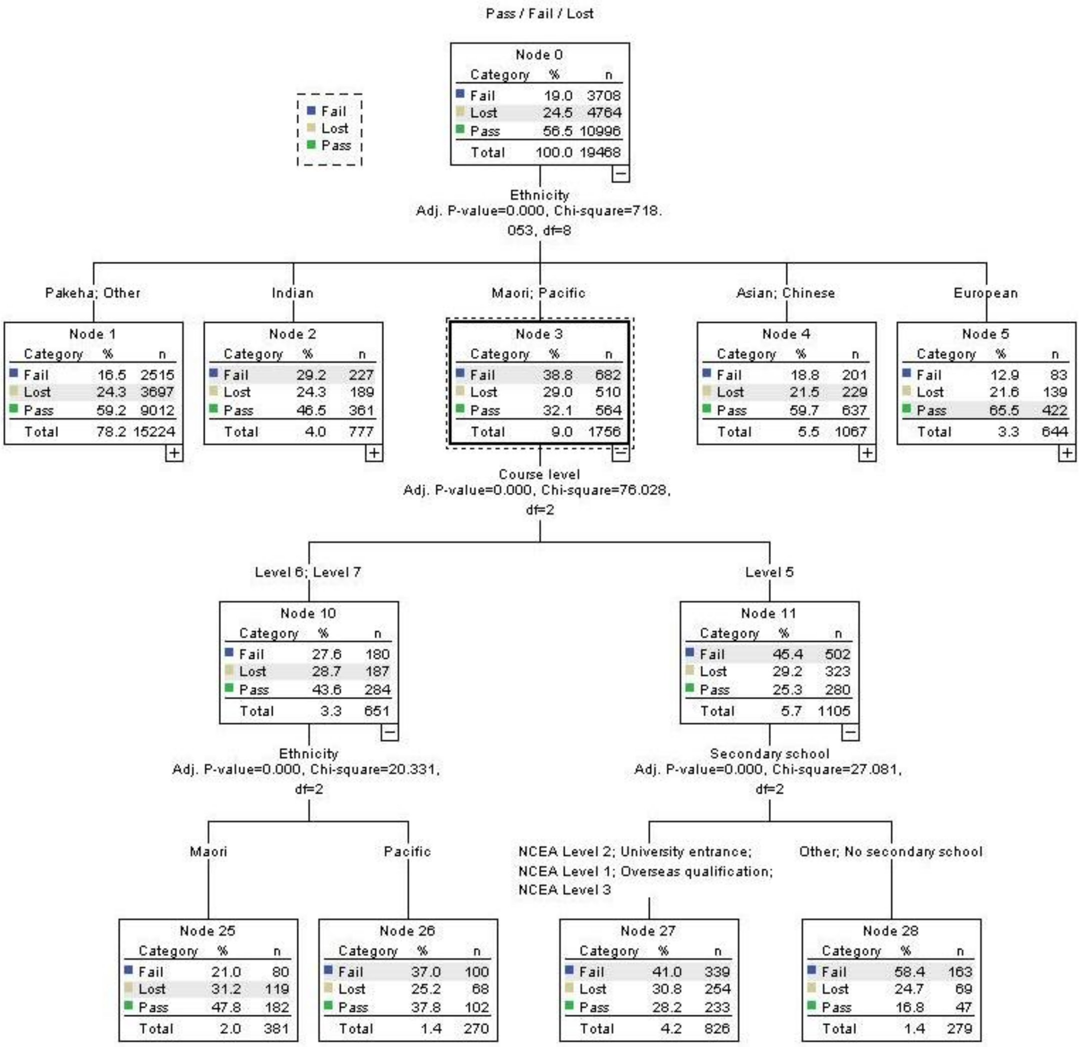


Figure 23: CHAID tree (Study outcome 1: Māori & Pacific Islanders)

Table 22: CHAID misclassification costs (Study outcome 1)

	Predicted		
Observed	Fail	Lost	Pass
Fail	0	1	2
Lost	1	0	2
Pass	1	1	0

Table 23: CHAID classification matrix (Study outcome 1)

	Predicted			
Observed	Fail	Lost	Pass	Percent correct
Fail	1636	1198	874	44.1%
Lost	1198	1939	1627	40.7%
Pass	1824	3703	5469	49.7%
Overall percentage	23.9%	35.1%	40.9	46.5%

Table 24: Rules for CHAID classification tree (Study outcome 1)

Node	Rule	Outcome	Probability
25	**IF** Ethnicity = “Māori” **OR** “Pacific Islander” **AND** Course level = “Level 6” **OR** “Level 7” **AND** Ethnicity = “Māori” **THEN**	Pass	0.478
26	**IF** Ethnicity = “Māori” **OR** “Pacific Islander” **AND** Course level = “Level 6” **OR** “Level 7” **AND** Ethnicity = “Pacific Islander” **THEN**	Pass	0.378
27	**IF** Ethnicity = “Māori” **OR** “Pacific Islander” **AND** Course level = “Level 5” **AND** Secondary school = “NCEA Level 1”**OR** “NCEA Level 2”**OR** “NCEA Level 3” **OR** “University entrance” **OR** “Overseas qualification” **THEN**	Fail	0.410
28	**IF** Ethnicity = “Māori” **OR** “Pacific Islander” **AND** Course level = “Level 5” **AND** Secondary school = “No secondary school” **OR** “Other” **THEN**	Fail	0.584

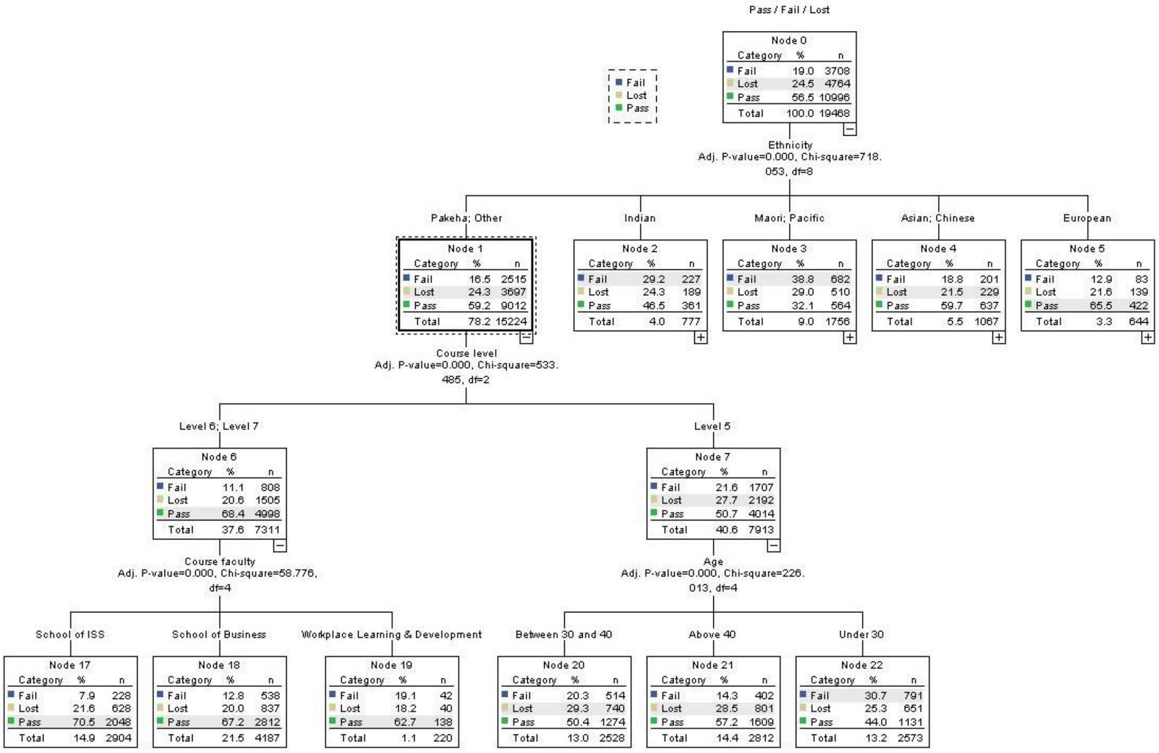


Figure 24: CHAID tree (Study outcome 1: Pakeha)

Table 25: Rules for CHAID tree (Study outcome 1: Pakeha)

Node	Rule	Outcome	Probability
17	**IF** Ethnicity = "Pakeha" **OR** "Others" **AND** Course level = "Level 6" **OR** "Level 7" **AND** Course faculty = "School of Information and Social Sciences" **THEN**	Pass	0.705
18	**IF** Ethnicity = "Pakeha" **OR** "Others" **AND** Course level = "Level 6" **OR** "Level 7" **AND** Course faculty = "School of Business" **THEN**	Pass	0.672
19	**IF** Ethnicity = "Pakeha" **OR** "Others" **AND** Course level = "Level 6" **OR** "Level 7" **AND** Course faculty = "Workplace Learning and Development" **THEN**	Fail	0.627
20	**IF** Ethnicity = "Pakeha" **OR** "Others" **AND** Course level = "Level 5" **AND** Age = "Between 30 and 40" **THEN**	Pass	0.504
21	**IF** Ethnicity = "Pakeha" **OR** "Others" **AND** Course level = "Level 5" **AND** Age = "Above 40" **THEN**	Pass	0.572
22	**IF** Ethnicity = "Pakeha" **OR** "Others" **AND** Course level = "Level 5" **AND** Age = "Under 30" **THEN**	Pass	0.440

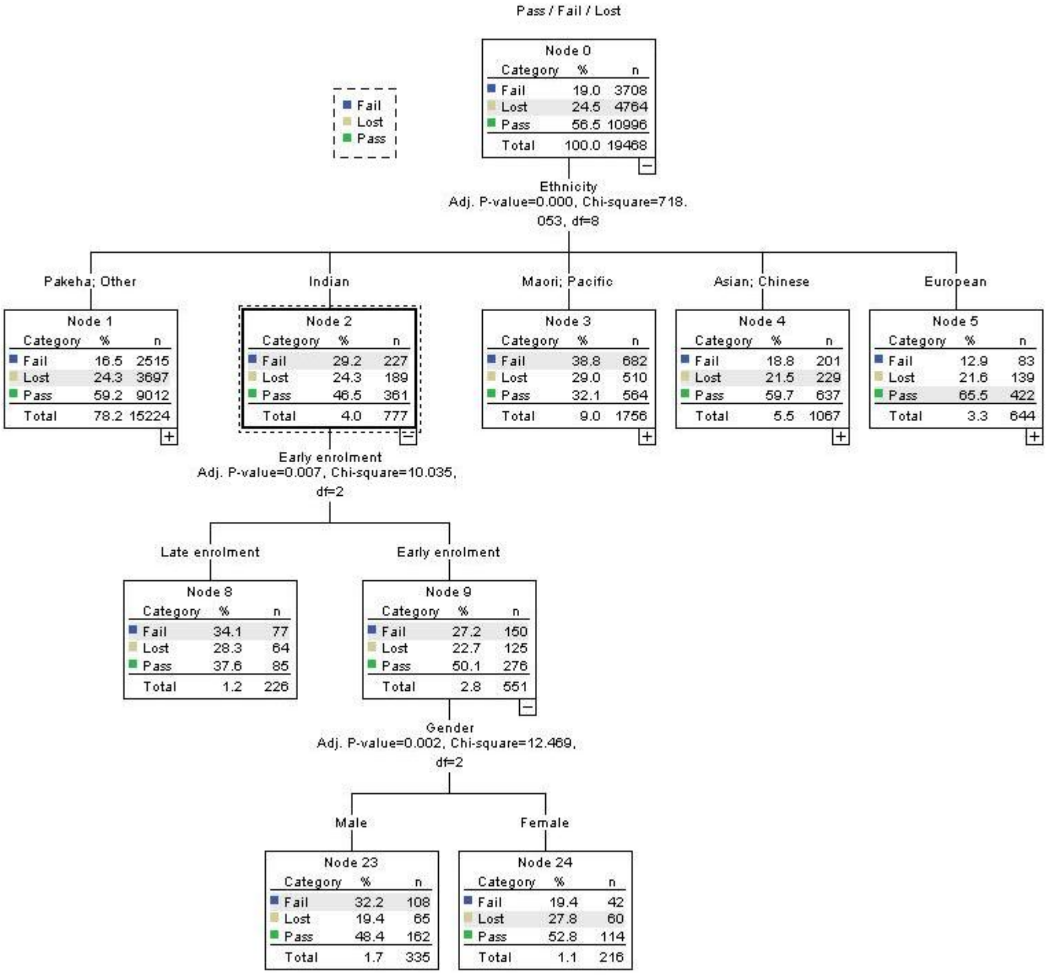


Figure 25: CHAID tree (Study outcome 1: Indian)

Table 26: Rules for CHAID tree (Study outcome 1: Indian)

Node	Rule	Outcome	Probability
8	**IF** Ethnicity = “Indian” **AND** Early enrolment = “Late” **THEN**	Pass	0.376
23	**IF** Ethnicity = “Indian” **AND** Early enrolment = “Early” **AND** Gender = “Male” **THEN**	Pass	0.484
24	**IF** Ethnicity = “Indian” **AND** Early enrolment = “Early” **AND** Gender = “Female” **THEN**	Pass	0.528

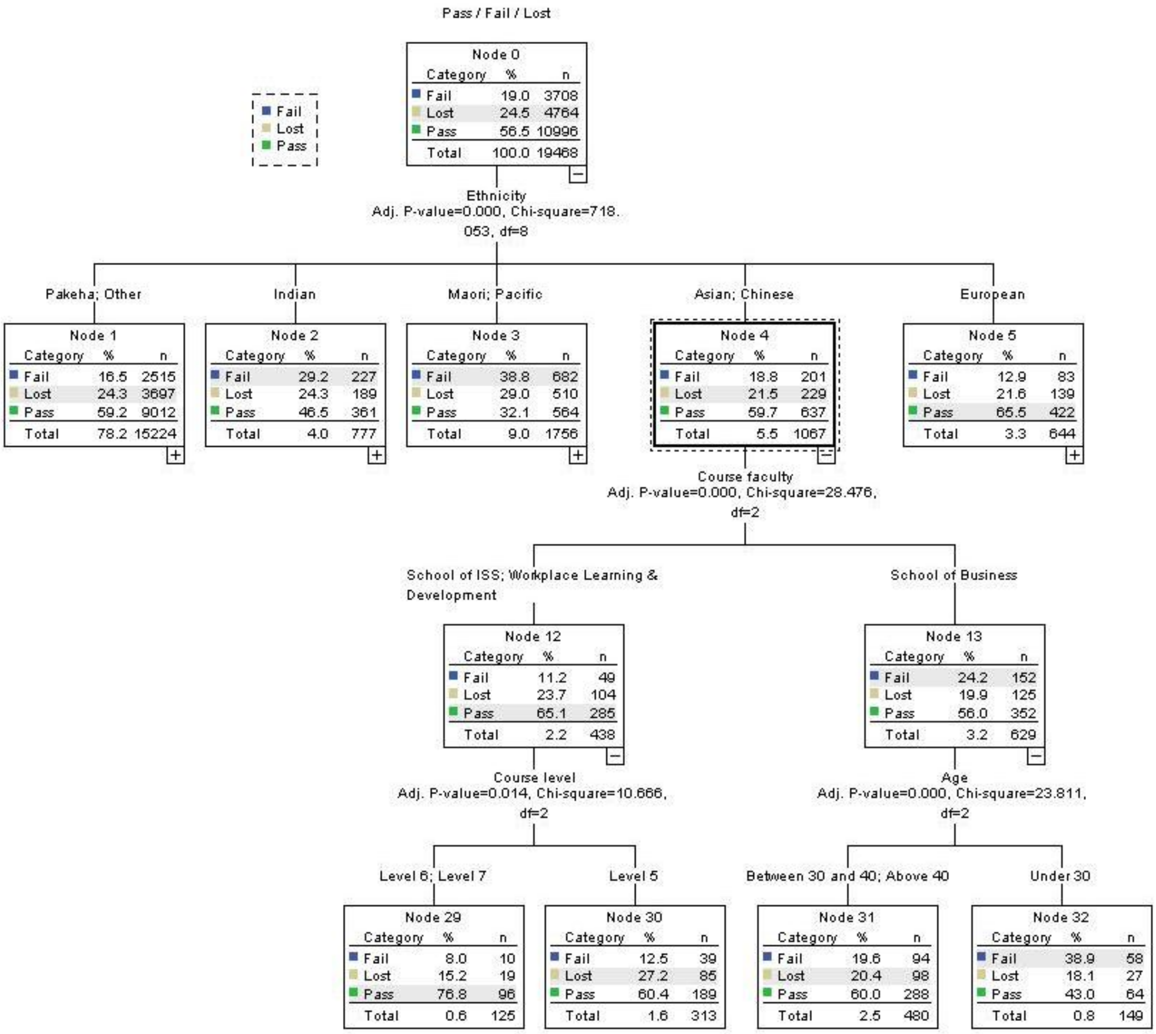


Figure 26: CHAID tree (Study outcome 1: Asian & Chinese)

Table 27: Rules for CHAID tree (Study outcome 1: Asian & Chinese)

Node	Rule	Outcome	Probability
29	**IF** Ethnicity = "Asian" **OR** "Chinese" **AND** Course faculty = "School of Information and Social Sciences" **OR** "Workplace Learning and Development" **AND** Course level = "Level 6" **OR** "Level 7" **THEN**	Pass	0.768
30	**IF** Ethnicity = "Asian" **OR** "Chinese" **AND** Course faculty = "School of Information and Social Sciences" **OR** "Workplace Learning and Development" **AND** Course level = "Level 5" **THEN**	Pass	0.604
31	**IF** Ethnicity = "Asian" **OR** "Chinese" **AND** Course faculty = "School of Business" **AND** Age = "Between 30 and 40" **OR** "Above 40" **THEN**	Pass	0.600
32	**IF** Ethnicity = "Asian" **OR** "Chinese" **AND** Course faculty = "School of Business" **AND** Age = "Under 30" **THEN**	Pass	0.430

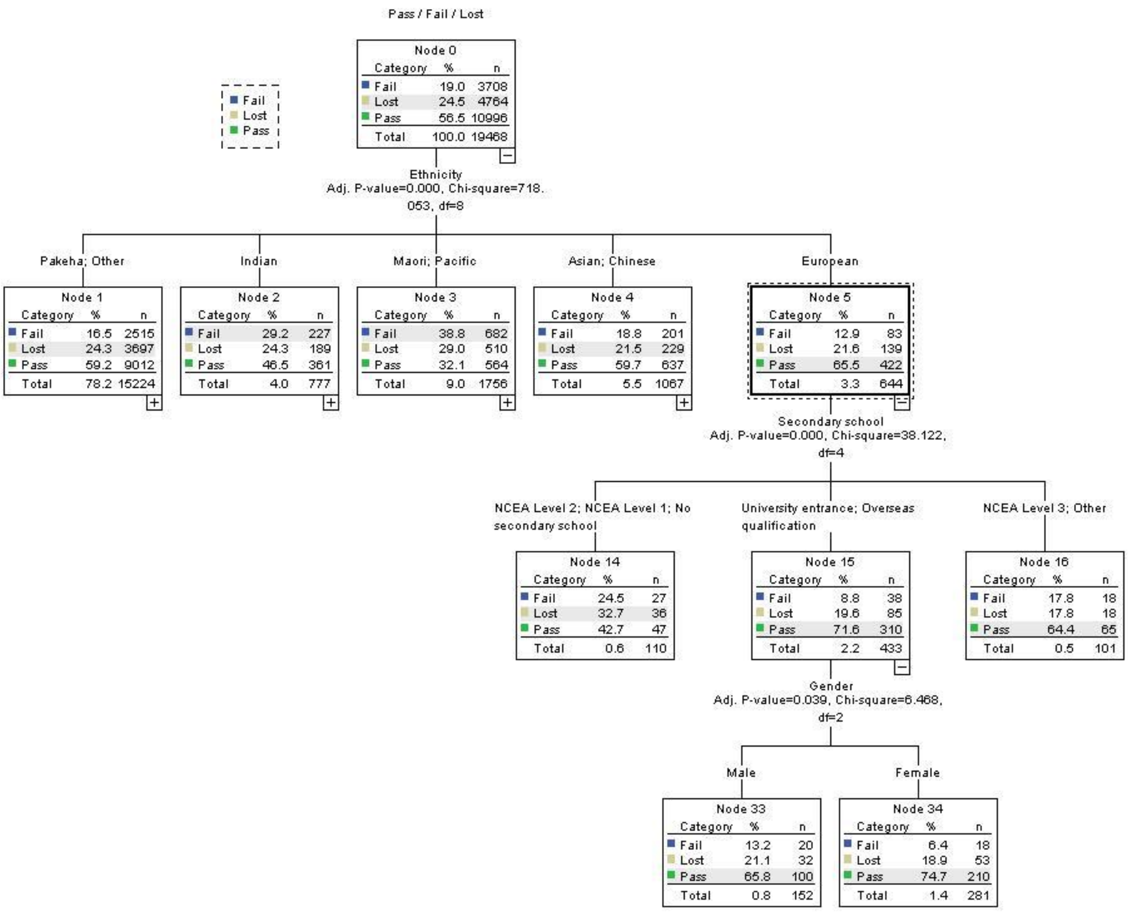


Figure 27: CHAID tree (Study outcome 1: European)

Table 28: Rules for CHAID tree (Study outcome 1: European)

Node	Rule	Outcome	Probability
14	**IF** Ethnicity = “European” **AND** Secondary school = “No secondary school” **OR** “NCEA Level 1” **OR** “NCEA Level 2” **THEN**	Pass	0.427
16	**IF** Ethnicity = “European” **AND** Secondary school = “NCEA Level 3” **OR** “Other” **THEN**	Pass	0.644
33	**IF** Ethnicity = “European” **AND** Secondary school = “University entrance” **OR** “Overseas qualification” **AND** Gender = “Male” **THEN**	Pass	0.658
34	**IF** Ethnicity = “European” **AND** Secondary school = “University entrance” **OR** “Overseas qualification” **AND** Gender = “Female” **THEN**	Pass	0.747

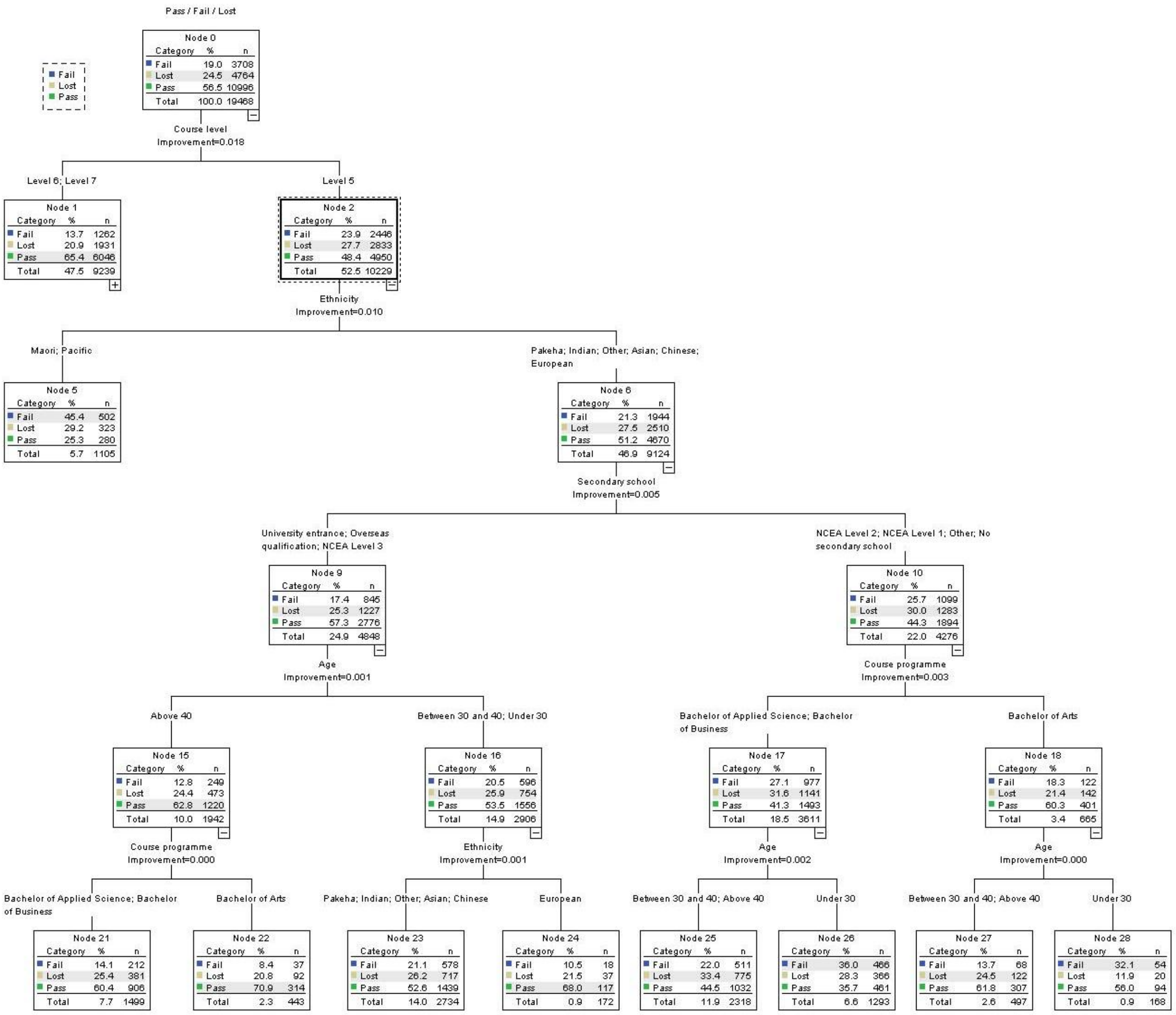


Figure 28: CART tree (Study outcome 1: Course level 5)

Table 29: CART misclassification costs (Study outcome 1)

	Predicted		
Observed	Fail	Lost	Pass
Fail	0	1	2
Lost	1	0	2
Pass	1	1	0

Table 30: CART classification matrix (Study outcome 1)

	Predicted			
Observed	Fail	Lost	Pass	Percent correct
Fail	1408	1394	906	38.0%
Lost	1064	2053	1647	43.1%
Pass	1509	3770	5717	52.0%
Overall percentage	20.4%	37.1%	42.5%	47.1%

Table 31: Rules for CART tree (Study outcome 1: Course level 5)

Node	Rule	Outcome	Probability
21	**IF** Course level = “Level 5” **AND** Ethnicity = “European” **OR** “Chinese” **OR** “Pakeha” **OR** “Asian” **OR** “Indian” **OR** “Others” **AND** Secondary school = “University entrance” **OR** “Overseas qualification” **OR** “NCEA Level 3” **AND** Age = “Above 40” **AND** Course programme = “Bachelor of Applied Science” **OR** “Bachelor of Business” **THEN**	Pass	0.604
22	**IF** Course level = “Level 5” **AND** Ethnicity = “European” **OR** “Chinese” **OR** “Pakeha” **OR** “Asian” **OR** “Indian” **OR** “Others” **AND** Secondary school = “University entrance” **OR** “Overseas qualification” **OR** “NCEA Level 3” **AND** Age = “Above 40” **AND** Course programme = “Bachelor of Arts” **THEN**	Pass	0.709
23	**IF** Course level = “Level 5” **AND** Ethnicity = “European” **OR** “Chinese” **OR** “Pakeha” **OR** “Asian” **OR** “Indian” **OR** “Others” **AND** Secondary school = “University entrance” **OR** “Overseas qualification” **OR** “NCEA Level 3” **AND** Age = “Under 30” **OR** “Between 30 and 40” **AND** Ethnicity = “Pakeha” **OR** “Indian” **OR** “Asian” **OR** “Others” **OR** “Chinese” **THEN**	Pass	0.526
24	**IF** Course level = “Level 5” **AND** Ethnicity = “European” **OR** “Chinese” **OR** “Pakeha” **OR** “Asian” **OR** “Indian” **OR** “Others” **AND** Secondary school = “University entrance” **OR** “Overseas qualification” **OR** “NCEA Level 3” **AND** Age = “Under 30” **OR** “Between 30 and 40” **AND** Ethnicity = “European” **THEN**	Pass	0.680
25	**IF** Course level = “Level 5” **AND** Ethnicity = “European” **OR** “Chinese” **OR** “Pakeha” **OR** “Asian” **OR** “Indian” **OR** “Others” **AND** Secondary school = “No secondary school” **OR** “NCEA Level 1” **OR** “NCEA Level 2” **OR** “Other” **AND** Course programme = “Bachelor of Applied Science” **OR** “Bachelor of Business” **AND** Age = “Above 40” **OR** “Between 30 and 40” **THEN**	Pass	0.445
26	**IF** Course level = “Level 5” **AND** Ethnicity = “European” **OR** “Chinese” **OR** “Pakeha” **OR** “Asian” **OR** “Indian” **OR** “Others” **AND** Secondary school = “No secondary school” **OR** “NCEA Level 1” **OR** “NCEA Level 2” **OR** “Other” **AND** Course programme = “Bachelor of Applied Science” **OR** “Bachelor of Business” **AND** Age = “Under 30” **THEN**	Pass	0.360
27	**IF** Course level = “Level 5” **AND** Ethnicity = “European” **OR** “Chinese” **OR** “Pakeha” **OR** “Asian” **OR** “Indian” **OR** “Others” **AND** Secondary school = “No secondary school” **OR** “NCEA Level 1” **OR** “NCEA Level 2” **OR** “Other” **AND** Course programme = “Bachelor of Arts” **AND** Age = “Above 40” **OR** “Between 30 and 40” **THEN**	Fail	0.618
28	**IF** Course level = “Level 5” **AND** Ethnicity = “European” **OR** “Chinese” **OR** “Pakeha” **OR** “Asian” **OR** “Indian” **OR** “Others” **AND** Secondary school = “No secondary school” **OR** “NCEA Level 1” **OR** “NCEA Level 2” **OR** “Other” **AND** Course programme = “Bachelor of Arts” **AND** Age = “Under 30” **THEN**	Pass	0.580

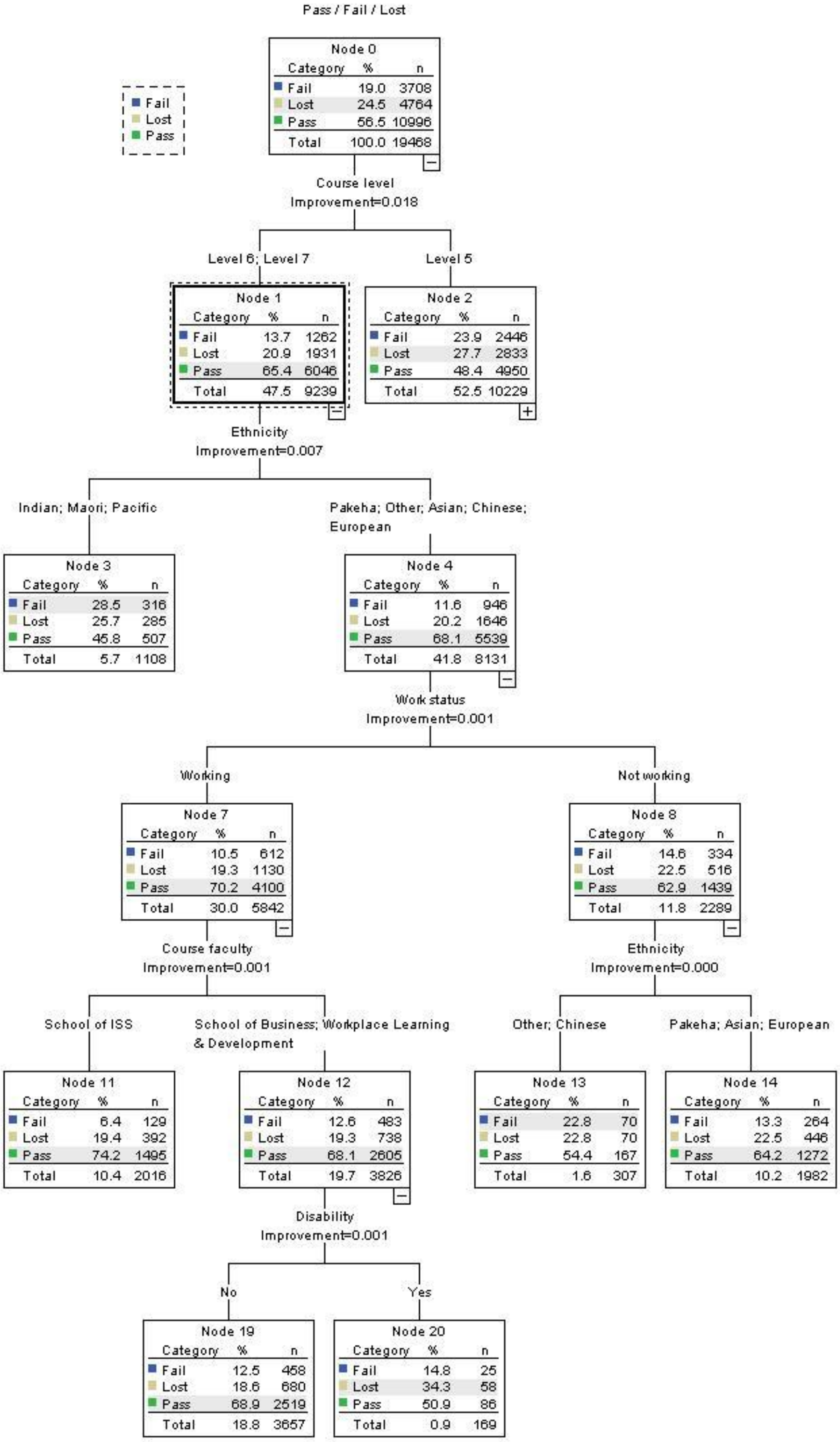


Figure 29: CART tree (Study outcome 1: Course level 6 & 7)

Table 32: Rules for CART tree (Study outcome 1: Course level 6 & 7)

Node	Rule	Outcome	Probability
3	**IF** Course level = “Level 6” **OR** “Level 7” **AND** Ethnicity “Māori” **OR** “Pacific Islander” **OR** “Indian” **AND** Secondary school = “University entrance” **OR** “Overseas qualification” **OR** “NCEA Level 3” **AND** Age = “Above 40” **AND** Course programme = “Bachelor of Applied Science” **OR** “Bachelor of Business” **THEN**	Pass	0.458
11	**IF** Course level = “Level 6” **OR** “Level 7” **AND** Ethnicity = “European” **OR** “Chinese” **OR** “Pakeha” **OR** “Asian” **OR** “Others” **AND** Work status = “Working” **AND** Course faculty = “School of Information and Social Sciences” **THEN**	Pass	0.742
13	**IF** Course level = “Level 6” **OR** “Level 7” **AND** Ethnicity = “European” **OR** “Chinese” **OR** “Pakeha” **OR** “Asian” **OR** “Others” **AND** Work status = “Not working” **AND** Ethnicity = “Chinese” **OR** “Others” **THEN**	Pass	0.544
14	**IF** Course level = “Level 6” **OR** “Level 7” **AND** Ethnicity = “European” **OR** “Chinese” **OR** “Pakeha” **OR** “Asian” **OR** “Others” **AND** Work status = “Not working” **AND** Ethnicity = “Pakeha” **OR** “European” **OR** “Asian” **THEN**	Pass	0.642
19	**IF** Course level = “Level 6” **OR** “Level 7” **AND** Ethnicity = “European” **OR** “Chinese” **OR** “Pakeha” **OR** “Asian” **OR** “Others” **AND** Work status = “Working” **AND** Course faculty = “School of Business” **OR** “Workplace Learning and Development” **AND** Disability = “No” **THEN**	Pass	0.689
20	**IF** Course level = “Level 6” **OR** “Level 7” **AND** Ethnicity = “European” **OR** “Chinese” **OR** “Pakeha” **OR** “Asian” **OR** “Others” **AND** Work status = “Working” **AND** Course faculty = “School of Business” **OR** “Workplace Learning and Development” **AND** Disability = “Yes” **THEN**	Pass	0.509

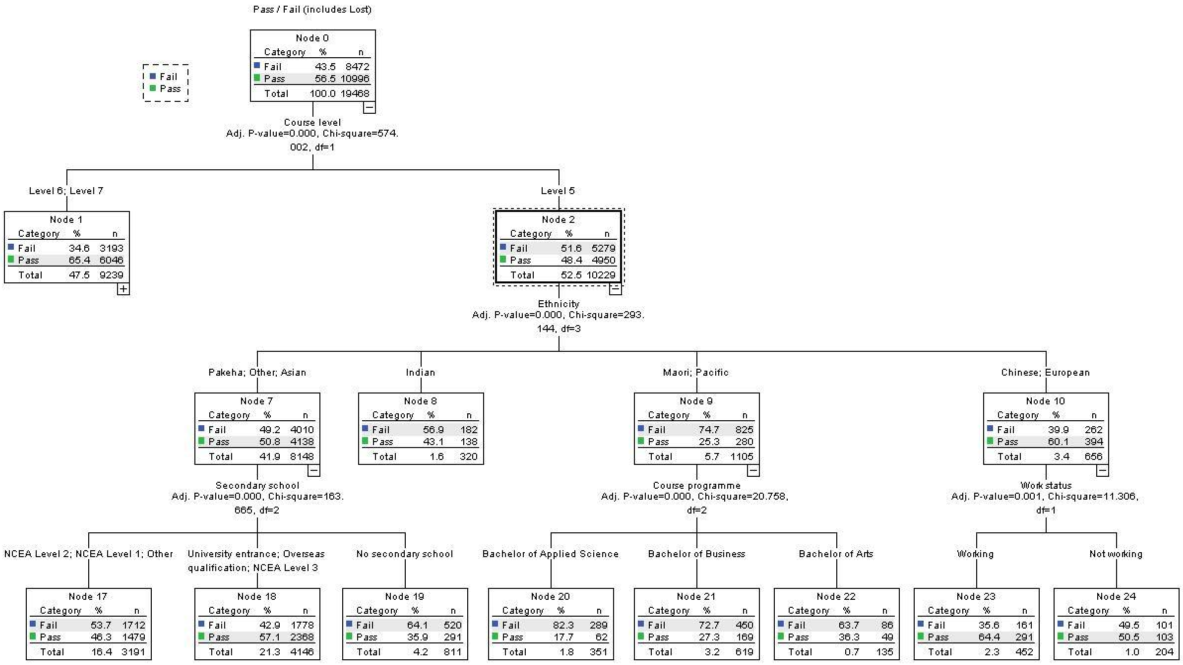


Figure 30: CHAID tree (Study outcome 2: Course level 5)

Table 33: Misclassification costs (Study outcome 2)

	Predicted	
Observed	Fail	Pass
Fail	0	1
Pass	1	0

Table 34: CHAID classification matrix (Study outcome 2)

	Predicted		
Observed	Fail	Pass	Percent correct
Fail	3499	4973	41.3%
Pass	2342	8654	78.7%
Overall percentage	30.0%	70.0%	62.4%

Table 35: Rules for CHAID tree (Study outcome 2: Course level 5)

Node	Rule	Outcome	Probability
8	**IF** Course level = “Level 5” **AND** Ethnicity = “Indian” **THEN**	Fail	0.569
17	**IF** Course level = “Level 5” **AND** Ethnicity = “Pakeha” **OR** “Asian” **OR** “Others” **AND** Secondary school = “NCEA Level 1” **OR** “NCEA Level 2” **OR** “Other” **THEN**	Fail	0.537
18	**IF** Course level = “Level 5” **AND** Ethnicity = “Pakeha” **OR** “Asian” **OR** “Others” **AND** Secondary school = “University entrance” **OR** “Overseas qualification” **OR** “NCEA Level 3” **THEN**	Pass	0.571
19	**IF** Course level = “Level 5” **AND** Ethnicity = “Pakeha” **OR** “Asian” **OR** “Others” **AND** Secondary school = “No secondary school” **THEN**	Fail	0.641
20	**IF** Course level = “Level 5” **AND** Ethnicity = “Māori” **OR** “Pacific Islander” **AND** Course programme = “Bachelor of Applied Science” **THEN**	Fail	0.823
21	**IF** Course level = “Level 5” **AND** Ethnicity = “Māori” **OR** “Pacific Islander” **AND** Course programme = “Bachelor of Business” **THEN**	Fail	0.727
22	**IF** Course level = “Level 5” **AND** Ethnicity = “Māori” **OR** “Pacific Islander” **AND** Course programme = “Bachelor of Arts” **THEN**	Fail	0.637
23	**IF** Course level = “Level 5” **AND** Ethnicity = “Chinese” **OR** “European” **AND** Work status = “Working” **THEN**	Pass	0.644
24	**IF** Course level = “Level 5” **AND** Ethnicity = “Chinese” **OR** “European” **AND** Work status = “Not working” **THEN**	Pass	0.505

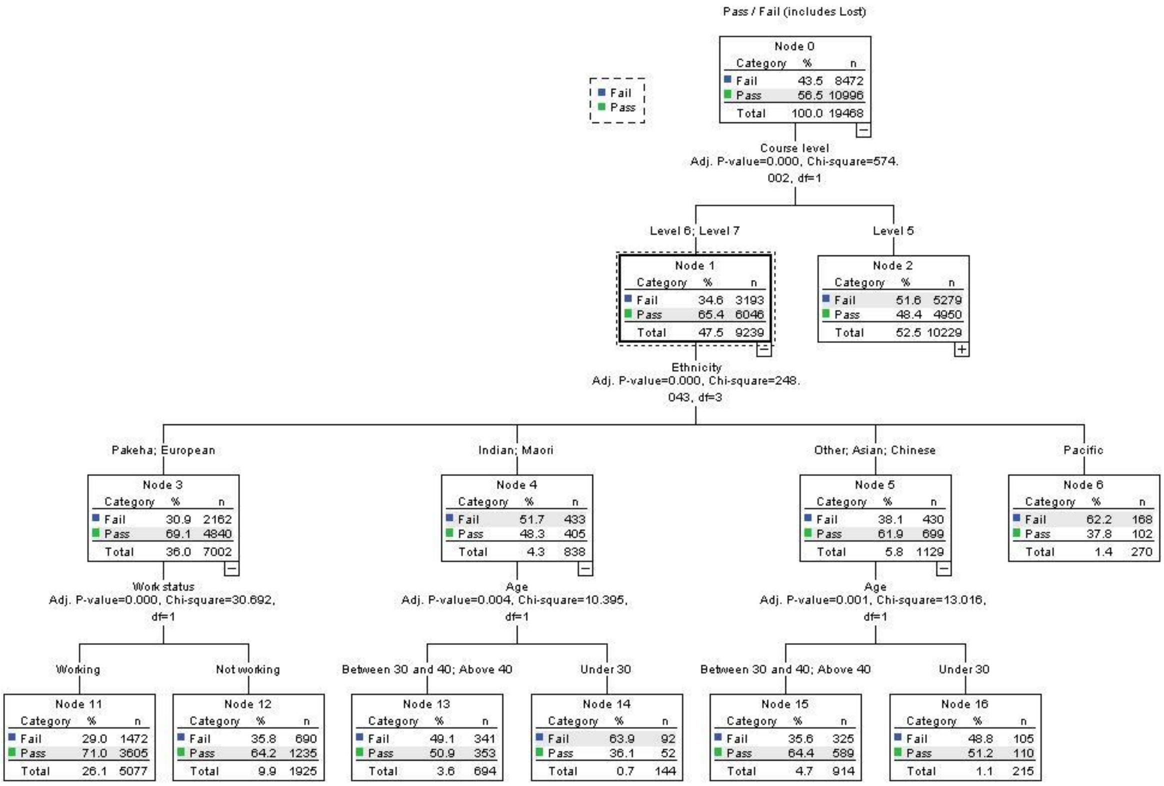


Figure 31: CHAID tree (Study outcome 2: Course level 6 & 7)

Table 36: Rules for CHAID tree (Study outcome 2: Course level 6 & 7)

Node	Rule	Outcome	Probability
6	**IF** Course level = "Level 6" **OR** "Level 7" **AND** Ethnicity = "Pacific" **THEN**	Fail	0.622
11	**IF** Course level = "Level 6" **OR** "Level 7" **AND** Ethnicity = "Pakeha" **OR** "European" **AND** Work status = "Working" **THEN**	Pass	0.710
12	**IF** Course level = "Level 6" **OR** "Level 7" **AND** Ethnicity = "Pakeha" **OR** "European" **AND** Work status = "Not working" **THEN**	Pass	0.642
13	**IF** Course level = "Level 6" **OR** "Level 7" **AND** Ethnicity = "Indian" **OR** "Māori" **AND** Age = "Between 30 and 40" **OR** "Above 40" **THEN**	Pass	0.509
14	**IF** Course level = "Level 6" **OR** "Level 7" **AND** Ethnicity = "Indian" **OR** "Māori" **AND** Age = "Under 30" **THEN**	Fail	0.639
15	**IF** Course level = "Level 6" **OR** "Level 7" **AND** Ethnicity = "Asian" **OR** "Chinese" **OR** "Others" **AND** Age = "Between 30 and 40" **OR** "Above 40" **THEN**	Pass	0.644
16	**IF** Course level = "Level 6" **OR** "Level 7" **AND** Ethnicity = "Asian" **OR** "Chinese" **OR** "Others" **AND** Age = "Under 30" **THEN**	Pass	0.512

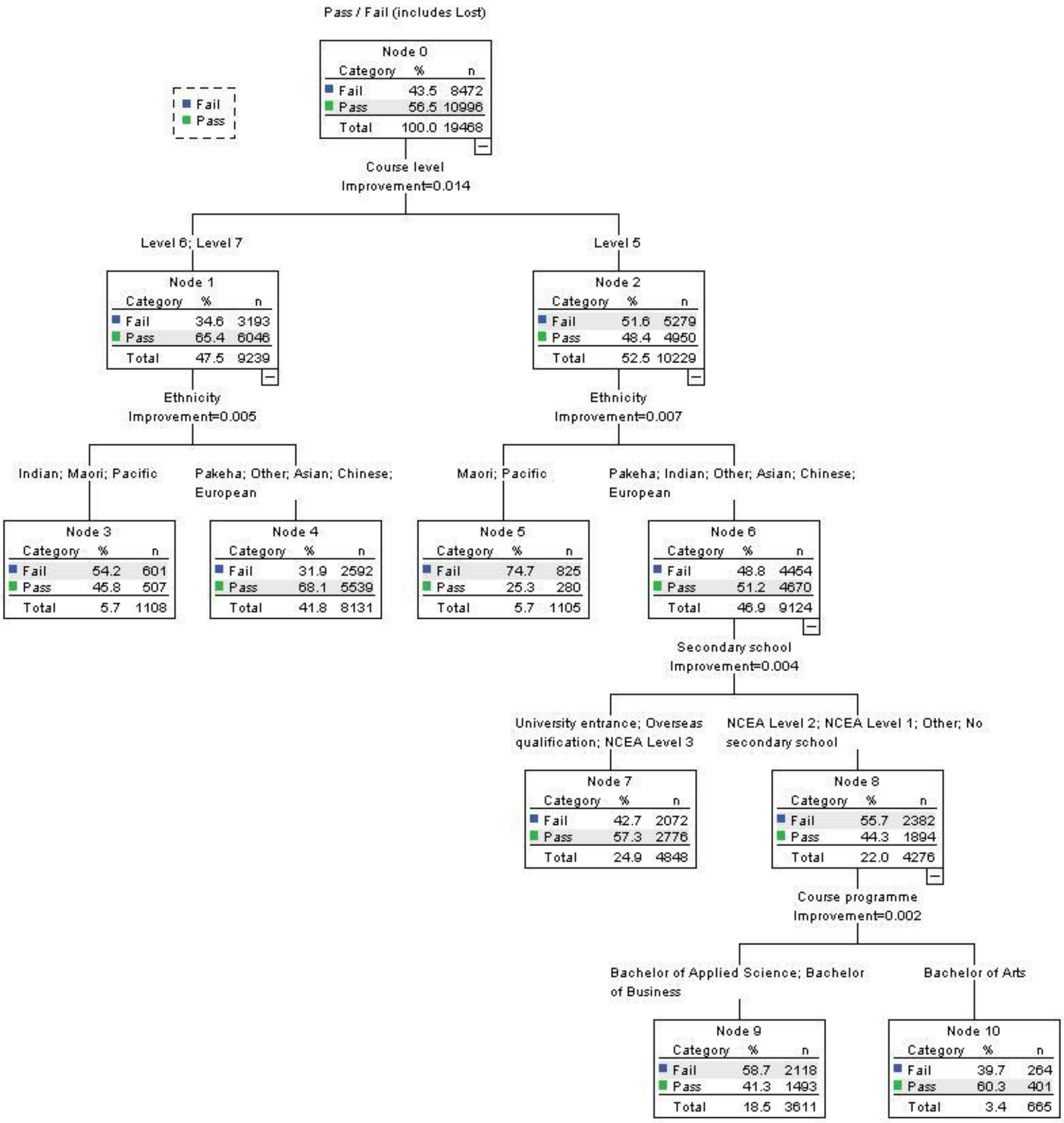


Figure 32: CART tree (Study outcome 2: Course level)

Table 37: CART classification matrix (Study outcome 2)

	Predicted		
Observed	Fail	Pass	Percent correct
Fail	3544	4928	41.8%
Pass	2280	8716	79.3%
Overall percentage	29.9%	70.1%	63.0%

Table 38: Rules for CART tree (Study outcome 2: Course level)

Node	Rule	Outcome	Probability
3	**IF** Course level = "Level 6" **OR** "Level 7" **AND** Ethnicity = "Indian" **OR** "Māori" **OR** "Pacific" **THEN**	Fail	0.542
4	**IF** Course level = "Level 6" **OR** "Level 7" **AND** Ethnicity = "Pakeha" **OR** "Asian" **OR** "European" **OR** "Chinese" **OR** "Others" **THEN**	Pass	0.681
5	**IF** Course level = "Level 5" **AND** Ethnicity = "Māori" **OR** "Pacific Islander" **THEN**	Fail	0.747
7	**IF** Course level = "Level 5" **AND** Ethnicity = "Pakeha" **OR** "Asian" **OR** "European" **OR** "Chinese" **OR** "Indian" **OR** "Others" **AND** Secondary school = "University entrance" **OR** "Overseas qualification" **OR** "NCEA Level 3" **THEN**	Pass	0.573
9	**IF** Course level = "Level 5" **AND** Ethnicity = "Pakeha" **OR** "Asian" **OR** "European" **OR** "Chinese" **OR** "Indian" **OR** "Others" **AND** Secondary school = "No secondary school" **OR** "NCEA Level 1" **OR** "NCEA Level 2" **OR** "Other" **AND** Course programme = "Bachelor of Applied Science" **OR** "Bachelor of Business" **THEN**	Fail	0.587
10	**IF** Course level = "Level 5" **AND** Ethnicity = "Pakeha" **OR** "Asian" **OR** "European" **OR** "Chinese" **OR** "Indian" **OR** "Others" **AND** Secondary school = "No secondary school" **OR** "NCEA Level 1" **OR** "NCEA Level 2" **OR** "Other" **AND** Course programme = "Bachelor of Arts" **THEN**	Pass	0.603

13. Appendix E: Logistic regression and discriminant analysis

Table 39: Binary logistic regression model (Study outcome 2)

Independent variable	Study outcome 2	
	Coefficient	Odds ratio
Intercept	-3.206	
Student demographics		
Gender	$0.058^{10\%}$	1.059
Age group		
Between 30 and 40	0.245	1.278
Above 40	0.410	1.506
Disability	0.385	1.469
Ethnic group		
European	1.450	4.262
Chinese	1.229	3.418
Pakeha	1.195	3.303
Asian	1.032	2.807
Other	0.987	2.683
Indian	0.573	1.773
Māori	0.317	1.373
Secondary school		
NCEA Level 1	0.352	1.421
NCEA Level 2	0.525	1.690
University Entrance	0.621	1.861
NCEA Level 3	0.867	2.380
Overseas qualification	0.725	2.065
Other	0.329	1.389
Work status	0.197	1.218
Early enrolment	0.105	1.111
Course characteristics		
Course faculty		
School of Infor. and Social Sciences	0.306	1.358
Workplace Learning and Develop.	0.564	1.758
Course programme		
OP7001 Bachelor of Business	0.194	1.214
OP7020 Bachelor of Arts	0.518	1.678
Course level		
Level 6	0.716	2.047
Level 7	0.661	1.937
Course block		
Semester 1	0.210	1.233
Semester 2	0.055^{ns}	1.056
Course offer type		
Distance	0.220	1.246
Blended	0.222	1.249
Number of observations	19468	
$-2 \log L$	24908.6	
Cox & Snell R^2	0.086	
Nagelkerke R^2	0.115	
Hosmer & Lemeshow test	$19.685^{5\%}$	
Overall % of correct classification	63.9%	

Note: Unless stated differently, all the coefficients are significant at less than the 1% level; *ns* stand for not significant.

Table 40: Polytomous logistic regression model (Study outcome 1)

Independent variable	Pass		Lost	
	Coefficient	Odds ratio	Coefficient	Odds ratio
Intercept	-3.708		-2.185	
Student demographics				
Gender	0.249	1.283	0.332	1.393
Age group				
Between 30 and 40	0.483	1.621	0.436	1.546
Above 40	0.806	2.239	0.690	1.995
Disability	0.357	1.429	-0.044^{ns}	0.957
Ethnic group				
European	1.913	6.773	0.858	2.358
Chinese	1.534	4.638	0.592	1.807
Pakeha	1.561	4.765	0.699	2.011
Asian	1.212	3.359	0.376$^{3\%}$	1.456
Other	1.405	4.077	0.787	2.198
Indian	0.683	1.981	0.226^{ns}	1.254
Māori	0.368	1.445	0.108^{ns}	1.114
Secondary school				
NCEA Level 1	0.480	1.616	0.246	1.279
NCEA Level 2	0.746	2.109	0.402	1.495
University Entrance	0.867	2.379	0.440	1.552
NCEA Level 3	1.174	3.236	0.552	1.736
Overseas qualification	0.949	2.582	0.403	1.497
Other	0.481	1.617	0.280$^{2\%}$	1.323
Work status	0.243	1.275	0.078^{ns}	1.081
Early enrolment	0.206	1.228	0.176	1.192
Course characteristics				
Course faculty				
School of Infor. and Social Sciences	0.540	1.717	0.399	1.491
Workplace Learning and Develop.	0.491	1.634	-0.166^{ns}	0.847
Course programme				
OP7001 Bachelor of Business	0.341	1.406	0.254	1.289
OP7020 Bachelor of Arts	0.524	1.689	0.008^{ns}	1.008
Course level				
Level 6	0.906	2.474	0.313	1.367
Level 7	0.706	2.027	0.079^{ns}	1.082
Course block				
Semester 1	0.313	1.367	0.184	1.202
Semester 2	0.182	1.199	0.222	1.249
Course offer type				
Distance	0.304	1.355	0.153$^{2\%}$	1.149
Blended	0.295	1.344	0.139^{ns}	1.166
$-2\log L$ initial / final	27070	2.4820		
Cox & Snell R^2	0.109			
Nagelkerke R^2	0.127			
McFadden	0.059			
Hosmer & Lemeshow test				
Overall % of correct classification	58.3%			

Note: Unless stated differently, all the coefficients are significant at less than the 1% level; *ns* stand for not significant.

Table 41: Discriminant function summary (Study outcome 1)

Variable	Standardised canonical coefficient		Structure coefficient		Wilks' Lambda	*F*
	1st function	2nd function	1st function	2nd function		
Student demographics						
Gender	-0.101	0.369	-0.083	0.396	0.998	23.8
Age group						
Under 30	0.377	-0.516	0.339	-0.350	0.985	143.3
Between 30 and 40	0.148	-0.124	0.022	0.097	1.000	1.5^{ns}
Above 40 [a]	-	-	-0.322	0.215	0.988	122.6
Disability	0.123	0.272	0.083	0.334	0.998	19.2
Ethnic group						
European	-0.110	-0.162	-0.105	-0.033	0.999	12.6
Chinese	-0.041	-0.176	-0.041	-0.097	1.000	$2.9^{6\%}$
Pakeha	-0.101	-0.325	-0.340	0.153	0.986	133.4
Asian	0.018	-0.187	0.001	-0.104	1.000	1.1^{ns}
Other [a]	-	-	-0.013	0.146	1.000	$2.4^{10\%}$
Indian	0.170	-0.195	0.155	-0.144	0.997	29.3
Māori	0.310	-0.308	0.376	-0.081	0.984	161.4
Pacific Islander	0.337	-0.232	0.348	-0.088	0.986	138.6
Secondary school						
No secondary school	0.162	-0.111	0.335	-0.035	0.987	128.0
NCEA Level 1	-0.013	-0.014	0.087	0.109	0.999	9.8
NCEA Level 2	-0.136	0.013	0.028	0.095	1.000	1.8^{ns}
University entrance	-0.204	-0.047	-0.140	0.034	0.998	22.5
NCEA Level 3	-0.294	-0.091	-0.147	-0.139	0.997	26.6
Overseas qualification	-0.229	-0.172	-0.121	-0.108	0.998	17.9
Other qualification [a]	-	-	0.130	0.046	0.998	19.4
Work status	-0.135	-0.083	-0.179	-0.101	0.996	37.4
Early enrolment	-0.089	0.109	-0.137	0.151	0.998	23.6
Course characteristics						
Course faculty						
School of Business	0.351	0.702	0.071	-0.225	0.999	10.8
School of ISS	0.057	0.985	-0.064	0.347	0.998	16.9
Workplace Learning [a]	-	-	-0.016	-0.319	0.999	10.6
Course programme						
Bachelor of Business	0.130	0.751	0.054	-0.031	1.000	$3.4^{3\%}$
Bachelor of Appl. Sci.	0.301	0.519	0.074	0.173	0.999	9.3
Bachelor of Arts [a]	-	-	-0.191	-0.202	0.995	45.7
Course level						
Level 5	0.455	0.569	0.515	0.356	0.969	314.5
Level 6	-0.068	0.189	-0.363	-0.145	0.985	151.5
Level 7 [a]	-	-	-0.232	-0.279	0.993	69.2
Course block						
Semester 1	-0.185	0.080	-0.161	-0.140	0.997	31.6
Semester 2	-0.094	0.293	0.016	0.191	1.000	$4.0^{2\%}$
Semester 3 [a]	-	-	0.216	-0.078	0.995	53.8
Course offer type						
Distance	0.001	0.005	-0.184	-0.030	0.996	38.5
Online	0.130	-0.029	0.175	-0.022	0.996	34.9
Blended [a]	-	-	0.061	0.068	1.000	4.68

Note: Unless stated differently, all the coefficients are significant at less than the 1% level; *ns* stand for not significant.

[a] This variable is not used in the analysis

Table 42: Discriminant analysis classification matrix
(Study outcome 1)

	Predicted			
Observed	Fail	Lost	Pass	Percent correct
Fail	843	42	2823	22.7%
Lost	539	88	4137	1.9%
Pass	512	63	10421	94.8%
Overall percentage	44.5%	45.6%	60.0%	58.3%

Table 43: Discriminant function summary (Study outcome 2)

Variable	Standardised canonical coefficient	Unstandardised canonical coefficient	Structure coefficient	Wilks Lambda	*F*
Intercept	-	-0.934			
Student demographics					
Gender	-0.042	-0.092	-0.026	1.000	1.24
Age group					
Under 30	0.290	0.671	0.288	0.992	153.33
Between 30 and 40	0.126	0.263	0.036	1.000	2.390
Above 40 [a]	-		-0.290	0.992	156.02
Disability	0.163	0.630	0.131	0.998	31.62
Ethnic group					
European	-0.133	-0.742	-0.109	0.999	22.19
Chinese	-0.068	-0.397	-0.055	1.000	$5.55^{2\%}$
Pakeha	-0.149	-0.337	-0.317	0.991	185.64
Asian [a]	-0.012	-0.077	-0.014	1.000	0.36^{ns}
Other [a]	-		0.008	1.000	0.11^{ns}
Indian	0.137	0.699	0.134	0.998	33.12
Māori	0.248	1.095	0.362	0.988	244.49
Pacific Islander	0.294	1.602	0.335	0.989	207.54
Secondary school					
No secondary school	0.142	0.529	0.330	0.990	201.19
NCEA Level 1	-0.015	-0.045	0.103	0.999	19.48
NCEA Level 2	-0.131	-0.330	0.042	1.000	$3.25^{7\%}$
University entrance	-0.208	-0.487	-0.135	0.998	33.85
NCEA Level 3	-0.302	-0.884	-0.167	0.997	51.64
Overseas qualification	-0.251	-0.656	-0.136	0.998	34.51
Other qualification [a]	-		0.136	0.998	34.38
Work status	-0.145	-0.315	-0.193	0.996	69.17
Early enrolment	-0.070	-0.172	-0.115	0.999	24.44
Course characteristics					
Course faculty					
School of Business	0.453	0.912	0.038	1.000	$2.71^{10\%}$
School of ISS	0.209	0.418	-0.014	1.000	0.38^{ns}
Workplace Learning [a]	-		-0.062	1.000	7.02
Course programme					
Bachelor of Business	0.244	0.495	0.049	1.000	$4.51^{3\%}$
Bachelor of Appl. Sci.	0.375	0.813	0.099	0.999	18.06
Bachelor of Arts [a]	-		-0.220	0.995	89.56
Course level					
Level 5	0.534	1.086	0.565	0.971	591.38
Level 6	-0.037	-0.084	-0.383	0.986	271.18
Level 7 [a]	-		-0.272	0.993	137.02
Course block					
Semester 1	-0.169	-0.342	-0.181	0.997	60.91
Semester 2	-0.047	-0.094	0.043	1.000	$3.43^{6\%}$
Semester 3 [a]	-		0.205	0.996	77.59
Course offer type					
Distance	0.001	0.003	-0.188	0.997	65.33
Online	0.123	0.366	0.172	0.997	54.56
Blended [a]	-		0.071	1.000	9.22

Note: Unless stated differently, all the coefficients are significant at less than the 1% level; *ns* stand for not significant.

[a] This variable is not used in the analysis

Table 44: Discriminant analysis classification matrix
(Study outcome 2)

	Predicted		
Observed	Fail	Pass	Percent correct
Fail	3753	4719	44.3%
Pass	2292	8704	79.2%
Overall percentage	62.1%	64.8%	64.0%