Early Prediction of Student Success: Mining Students Enrolment Data

Zlatko J. Kovačić
Open Polytechnic, Wellington, New Zealand

Zlatko.Kovacic@openpolytechnic.ac.nz

Abstract

This paper explores the socio-demographic variables (age, gender, ethnicity, education, work status, and disability) and study environment (course programme and course block), that may influence persistence or dropout of students at the Open Polytechnic of New Zealand. We examine to what extent these factors, i.e. enrolment data help us in pre-identifying successful and unsuccessful students.

The data stored in the Open Polytechnic student management system from 2006 to 2009, covering over 450 students who enrolled to *71150 Information Systems* course was used to perform a quantitative analysis of study outcome. Based on a data mining techniques (such as feature selection and classification trees), the most important factors for student success and a profile of the typical successful and unsuccessful students are identified.

The empirical results show the following: (i) the most important factors separating successful from unsuccessful students are: ethnicity, course programme and course block; (ii) among classification tree growing methods Classification and Regression Tree (CART) was the most successful in growing the tree with an overall percentage of correct classification of 60.5%; and (iii) both the risk estimated by the cross-validation and the gain diagram suggests that all trees, based only on enrolment data are not quite good in separating successful from unsuccessful students. The implications of these results for academic and administrative staff are discussed.

Keywords: Study outcome, persistence, dropout, distance education, classification tree

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 criteria for government funding in the tertiary education environment in New Zealand is the level of 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 most vulnerable students to low student retention at all institutions of higher education are the 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 on time. 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 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.

The 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, more 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, background characteristics taken alone as predictors of dropout have not performed well in case of traditional students (regular, full-time students). However, the background information was significant in case of non-traditional students (distance/open education) where social integration and institutional commitment are not central in the student experience.

At the time of enrolment in the Open Polytechnic of New Zealand, 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 student.

Research Objectives

The main objective of this study is to explore factors that may impact the study outcome in the *Information Systems* course at the Open Polytechnic. The *Information Systems* course is a core course for those majoring in IT and for most students an entry point, i.e. the first course they are taking with the Open Polytechnic. This issue have not been examined so far for Open Polytechnic and this paper attempts to fill the gap. More specifically the enrolment data were used to achieve the following objectives:

- Build models for early prediction of study outcome using the 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)

Before answering these questions a brief overview of the previous empirical studies of study outcome prediction based on enrolment data is given. Based on the literature review determinants of study outcome were identified and discussed. 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.

Literature Review

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. Traditionally, from the methodological point of view, statistical models such as logistic regression and discriminant analysis were used most frequently in retention studies to identify factors and their contributions to the student dropout. There are also other, less frequently used models such as survival or failure-time analysis (Murtaugh, Burns & Schuster, 1999), 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. Romero & Ventura (2007) provided a survey of educational data mining from 1995-2005 and Baker & Yacef (2009) extended their survey covering the latest development until 2009. There are an increasing number of data mining applications in education, from enrollment management, graduation, academic performance, gifted education, web-based education, retention and other areas (Nandeshwar & Chandhari, 2009). In this section we will review only 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 student will have to deal with during his/her study.

Woodman (2001) found for courses in the mathematics and computing faculty at the Open University in UK, by using the binary logistic regression, that the most significant factors to whether students passed, failed or dropped out, were marks for the first assignment, the number of maths courses passed in the previous two years, the course level, the points the course is worth and the occupation group of the student. This was the most parsimonious model, but in the 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 the logistic regression is that in large samples any difference, may lead to conclusion that the factor is significant when in fact that 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 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 the supervised machine learning algorithms (decision trees, artificial neural networks, naïve Bayes classifier, instance-based learning, logistic regression and support vector machines) to predict student's performance at the Hellenic Open University. When only the demographic variables were used the prediction accuracy varied from 58.84% (when using 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 independ-

ence, 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' was the most important factor contributing to retention. Gender and ethnic origin were not identified as significant.

Al-Radaideh, Al-Shawakfa & Al-Najjar (2006) were using the classification trees to predict the final grade 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.

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 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 the 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 age has a significant positive effect. The variable was entered in a quadratic form to allow the effect of age to have 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 student should be constructed for each academic 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 the course / diploma /degree in a timely manner despite the 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, students population and classification methods used. Even when the background informa-

tion was significantly related to the academic performance, the prediction accuracy was pretty low with an overall accuracy around 60%.

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 large, noisy, and messy data set, which can lead to actions to increase benefit in some form (diagnosis, profit, detection, etc.)"

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, modeling, evaluation and deployment. The feedback from deployment to data and business understanding illustrates the iterative nature of a data mining process.
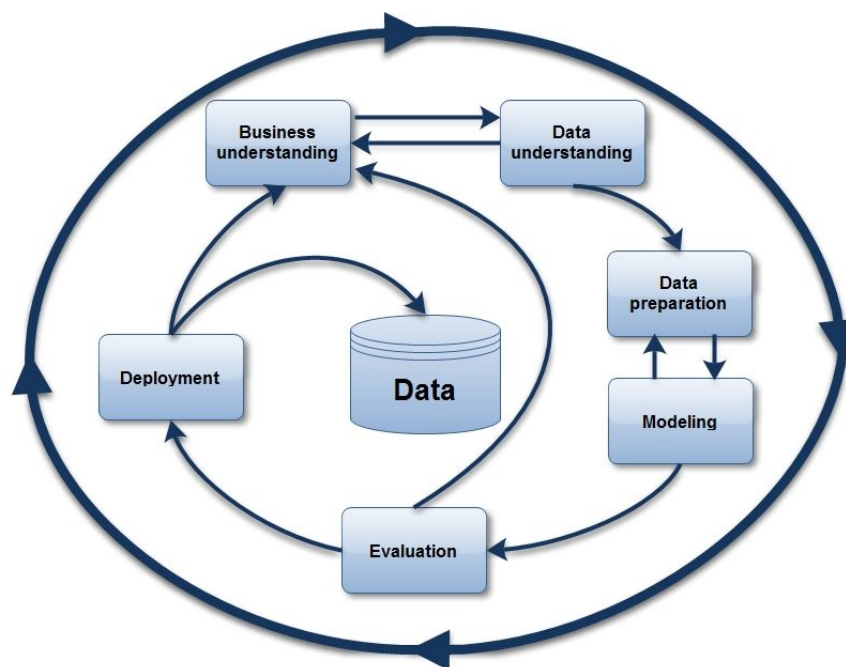

Figure 1: Modified CRISP-DM Model Version 1
(Adopted from Nisbet, Elder & Miner, 2009)

Business Understanding

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

Because we are planning to increase completion rate on the *Information Systems* course understanding its students population and patterns in the pre-enrolment data becomes necessary before we start developing a predictive model. In this phase we are coming up with the following questions: what is the profile of a student who successfully completes this course? Can the successful

vs. unsuccessful student be distinguished in terms of demographic features (such as gender, age or ethnic origin) or study environment (such as course programme, faculty or course block)? Depending on the answers to these questions we consider the methods and approaches that can be adopted to increase the completion rate.

Data Understanding

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 in the data.

The enrolment form asks students to enter the following information: demographic (gender, date of birth, ethnic origins, disability, and work status), academic (secondary school qualification, course programme, course faculty, and course block), and contact details. Once the data from the enrolment form is entered into Integrator and the Enrolment Section process the application, the enrolment date is recorded and student becomes enrolled on the course.

Data Preparation

In this phase the data are put into a form suitable for the modeling phase. If required some selected variables are combined, transformed or used to create a new variables. For example, enrolment date and the course block start date were used to generate a variable labeled as "early enrolment". Any data excluded from the data set is documented and their removal explained. For example, only a few students enrolled to Bachelor of Arts programme were on the *Information Systems* course. They were removed from further analysis for the reason explained later in Data and Methodology section. Data are cleaned for any duplication of records. For example, in case of *Information Systems* course, the course code changed in the past. If student enrolled in 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 two possible outcomes (labeled as *Pass* and *Fail*) indicates whether students successfully completed the course or failed the course due to voluntary transfer/withdrawal or academic withdrawal or simply because not fulfilling course pass requirements.

Modeling

In this phase we choose 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 analytical and managerial standpoint. In this phase we decided to use classification tree models with four different tree growing criteria.

Evaluation

The final models from the previous phase are then applied on a testing, i.e. a validation data set with the aim of assessing their predictive accuracy and consistency with the results obtained for the training data set. This phase involves an iterative process of fitting different versions of models to training and testing data set, each time evaluating their predictive performance.

Deployment

Once we decided on the final model we can apply it to current data not used during the modeling and evaluation phase. This process is known as *scoring*. The model results are used to address the

issues identified in the business understanding phase. The results should be presented in the user-friendly format and prepared for use by administrative staff. The final model should retain the highest predictive accuracy and if it is to be continuously used it should be regularly updated, particularly if some organizational changes occur or if new factors are brought in. For example, if the new information about financial support is added to the student record or enrolment form, then the model should consider the new factor that might be relevant for study outcome.

Data and Methodology

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

Data preparation

Variables definition and their domains are presented in Table 1.

Until 2007 the *Information Systems* course was offered as a distance and online course with a separate course codes. From 2007 this course is offered exclusively as an online course. Therefore for the pre-2007 data, the course ID was changed and both, distance and online occurrences of the course were coded as one course.

From the initial dataset all students granted cross-credit or credit were excluded because they didn't actually study this course. The courses they have previously completed were recognized and credited for this course. We also removed 6 students enrolled under Bachelor of Arts programme. Such a small number of students in this category do not allow generalization of the results for this particular category. The total number of data was reduced to 453.

We needed to clarify the definition of study outcome that we used in our analysis. We considered only two possible outcomes, labeled as: *Pass* and *Fail*. Students labeled *Pass* successfully completed the course. Students labeled *Fail* transfered or withdrew from the course voluntarily or academic registry withdrew them for not completing the in-course assessments. Those students who stayed on the course until the end of the course but scored less than the course pass mark were also allocated into this category.

Because of the data set size (only 453 students) numeric continuous variable such as age was converted into a categorical variable with only three age groups. In Ethnicity we combined Maori and Pacific Island students because of two reasons: they were found to be no different in preliminary bivariate analysis and combined together constitute a small proportion of the data (less than 10%). Combining them into one ethnic group helps with model parsimony. The Secondary school variable combines all students with no secondary school up to NCEA Level 2 on the New Zealand National Qualifications Framework into one group.

In data mining variables are also known as features, predictors or attributes. Though in some areas of machine learning, a distinction is made between a variable and a feature (Nisbet, Elder & Miner, 2009) we will use them interchangeably.

Though the software applications skills (spreadsheet and database in particular) are very important on this course we do not take into account the different skill levels present on the *Information Systems* course. Since the students enrolling on this course have different backgrounds and levels of interest in computing, we would expect that skill level has also a significant impact on the study outcome. However, the information about their Office skills level is not available in the moment of enrolment so we did not include this factor in our analysis.

Table 1: Description of variables and their domains

Variable	Description (Domain)
	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)
Ethnicity	Student's ethnic group (nominal: Pakeha, Maori & Pacific Islanders or Others)
Disability	Student has a disability (binary: yes or no)
Secondary school	Student's highest level of achievement from a secondary school on the New Zealand National Qualifications Framework (nominal: No secondary qualification, NCEA1, NCEA2, University entrance, NCEA3, Overseas or Other)
Work status	Student is working (binary: yes or no)
Early enrolment	Student enrolled for the first time in the course before start of the course (binary: yes or no)
	Study environment
Course programme	Programme (nominal: Bachelor of Business or Bachelor of Applied Science)
Course block	Semester in which a course is offered (Semester 1, Semester 2 or Semester 3)
	Dependent variable
Study outcome	Study outcome (nominal: Pass – successful completion, Fail – unsuccessful completion includes also withdrawals, academic withdrawals and transfers)

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 modeling study outcome. The third type of data mining approach, i.e. predictive data mining is conducted by using four different classification trees. Finally, a comparison between these classification tree models was conducted to determine the best model for the dataset. Data were analyzed using SPSS 17 and Statistica 8.

We decided to use the classification tree models because of some advantages they may have 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.

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. Finally, the classification tree results for four different growing methods are presented.

Summary Statistics

As part of the data understanding phase we carried out the cross-tabulation for each variable and the study outcome after preparing and cleaning the data. The Table 2 reports the results. Based on the results shown majority of *Information Systems* students are female (over 63%). However, percentage of female students who successfully complete the course are higher (65%) which suggests that female students are more likely to pass the course than their male counterpart. When it comes to age over 68% of students are above 30 with the age group between 30 and 40 being majority. This age group is also more likely to fail the course because their 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%).

Table 2: Descriptive statistics (percentage) – Study outcome (453 students)

Variable	Domain	Count	Total	Pass	Fail
Gender	Female	286	63.1	65.0	61.5
	Male	167	36.9	35.0	38.5
Age	Under 30	136	30.0	30.4	29.7
	Between 30 and 40	175	38.6	37.4	39.7
	Above 40	142	31.3	32.2	30.5
Disability	Yes	19	4.2	3.3	5.0
	No	434	95.8	96.7	95.0
Ethnicity	Pakeha	318	70.2	75.7	65.3
	Maori & Pacific Islanders	41	9.1	2.8	14.6
	Others	94	20.8	21.5	20.1
Secondary school	No secondary school / NCEA Level 1 or 2	183	40.4	36.9	43.5
	University Entrance / NCEA Level 3	163	36.0	38.3	33.9
	Overseas Qualification or Other	107	23.6	24.8	22.6
Work status	Yes	351	77.5	78.0	77.0
	No	102	22.5	22.0	23.0
Early enrolment	Yes	317	70.0	72.4	67.8
	No	136	30.0	27.6	32.2
Course programme	Bachelor of Business	305	67.3	73.8	61.5
	Bachelor of Applied Sciences	148	32.7	26.2	38.5
Course block	First semester	139	30.7	31.3	30.1
	Second semester	201	44.4	48.6	40.6
	Third semester	113	24.9	20.1	29.3

Disability was shown to be a disadvantage for *Information Systems* students. Students with it are more likely to fail than those without it. There are huge differences in percentage of students who

successfully completed the course depending on their ethnic origin. Though Maori and Pacific Islanders make 9.1% of all students on this course their participation is significantly lower in the "Pass" subpopulation (i.e. 2.8%) and higher in the "Fail" subpopulation (14.6%). Based on these results we can say that students with this ethnic origin are identified as students "at- risk". Further methods of data mining will confirm this statement.

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

We are using 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 enroll well before the enrolment closing date. The opposite category (late comers) participates with 30% in 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%.

One third of students on this course 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 Business programme. Finally, students studying this course in the summer semester are more likely to fail than those studying in the second and first semester.

Feature Selection

The number of predictor variables is not so large and we don't have to select the 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 with dependent or outcome variable. 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.
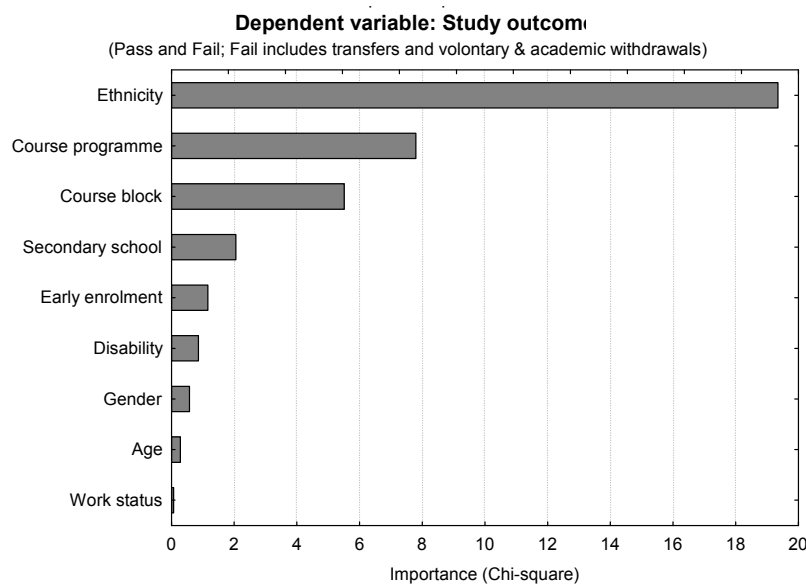

Dependent variable: Study outcome
(Pass and Fail; Fail includes transfers and volontary & academic withdrawals)

Figure 2: Importance plot for predictors

Results of feature selection are presented in Figure 2 on the importance plot and also in Table 3. The top three predictors for the study outcome are: ethnic origin of students, course programme they are enrolled on and course block, i.e. semester in which they are study.

Table 3: Best predictors for dependent variable

Variable	Chi-square	*P*-value
Ethnicity	19.35	0.00006
Course programme	7.80	0.00523
Course block	5.51	0.06354
Secondary school	2.06	0.35748
Early enrolment	1.16	0.28131
Disability	0.86	0.35363
Gender	0.58	0.44774
Age	0.28	0.86750
Work status	0.07	0.78940

From Table 3, *P*-values from the last column only the first three chi-square values are significant at 10% level. Though the results of the feature selection suggested continuing analysis with only the subset of predictors, which includes ethnicity, course programme and course block, we have included all available predictors in our classification tree analysis. We follow an 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.

Classification Trees

The objective of an analysis based on a classification tree is to identify factors that contribute the most to 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 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).

The classification trees for study outcome are given in Figure 2 and 4. In each tree node the number of successful students (4th line, last column) and unsuccessful students (3rd line, last column) is given, as well as the percentages for each category (2nd column) and 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, 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 their insignificance detected in the feature selection section.

CHAID, exhaustive CHAID and QUEST

Three classification tree growing methods, namely: CHAID, exhaustive CHAID and QUEST generated exactly the same tree structure presented in Figure 3. It shows that only 2 variables were used to construct the tree: (1) ethnicity and (2) course programme. All the other student de-

mographics variables were used but not included in the final model. We could change the stopping criteria to allow further growing of the tree, but that would 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 more likely poor performance when used on a new data set. This phenomenon is known as overfitting the tree.

The largest successful group (i.e. students who successfully completed the course) consists of 274 (60.5%) students (Node 3). Ethnic origin of students in this group is either Pakeha or other ethnic groups (excluding Maori and Pacific Islands students). Students in this group opted for the Bachelor of Business programme. The largest unsuccessful group (i.e. students who were unsuccessful) contains 138 students (30.5% of all participants) (Node 4). They belong to either Pakeha or other ethnic groups (excluding Maori and Pacific Islands students). The next largest group considered also as unsuccessful students, contains 41, i.e. 9.1% of all students, where 75.4% of them are unsuccessful (Node 1). They are described as Maori and Pacific Islands students.
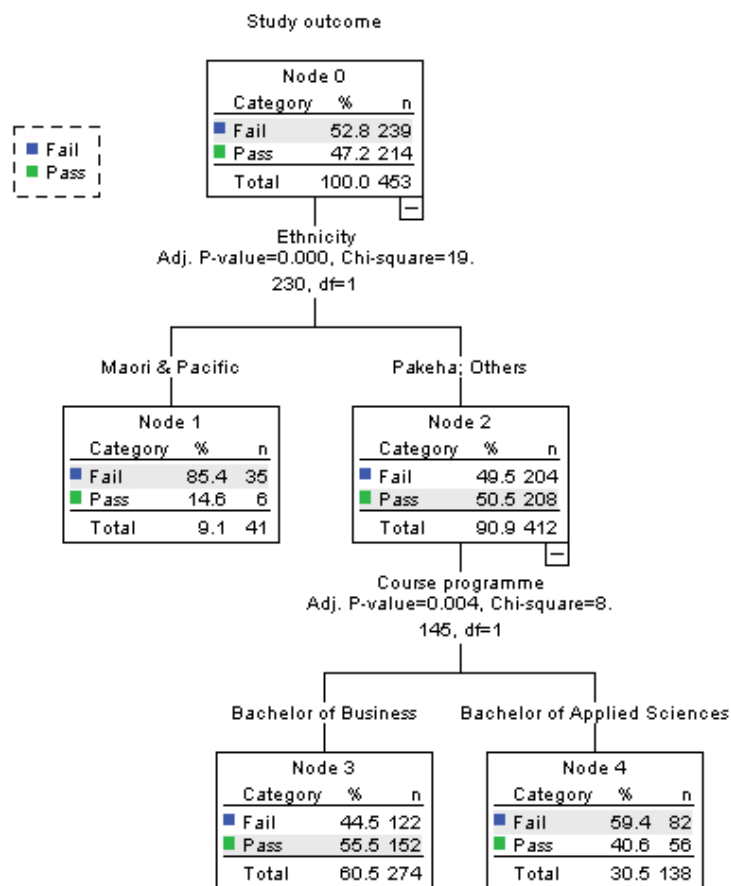

Figure 3: CHAID, exhaustive CHAID and QUEST classification tree

The overall percentage of correct classification for the study outcome is only 59.4% (Table 4). This percentage of correct classification was achieved with 2 variables only.

The cross-validation estimate of the risk is 0.406 indicates that the category predicted by the model (successful or unsuccessful student) is wrong for 40.6% of the cases. So the risk of misclassifying a student is approximately 41%. This result is consistent with the results in the CHAID classification matrix (Table 4). The Overall percentage shows that the model only classified correctly 59% of students.

Table 4: CHAID classification matrix

Observed	Predicted		
	Fail	Pass	Percent correct
Fail	117	122	49.0%
Pass	62	152	71.0%
Overall percentage	65.4%	55.5%	59.4%

With large numbers of false positives (122) and few false negatives (62), the CHAID model is in itself poor at identifying an unsuccessful student (positive predictive value is only 49%). It will, however, pick up 65.4% of all unsuccessful students (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 71%) and correctly identifies 55.5% of those who pass the course (the specificity).

The classification matrix indicates also another problem with the model. For unsuccessful students it predicts failure for only 49% of them, which means that 51% of unsuccessful students are inaccurately classified with the successful students. Practical consequence of this misclassification is that these students would not received additional learning support provided to the students "at risk", simply because they will be classified by the model among successful students. This feature of the model is more critical than misclassification of the successful students among unsuccessful students (29% of successful students belong to this category). In this case these students may receive additional learning support even though they don't need it. One option to increase 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. In this case the increased cost for misclassification of unsuccessful to the successful group of students decreased significantly both remaining percentage of correct classification which was not compensated in equivalent increase in the initial 49%.
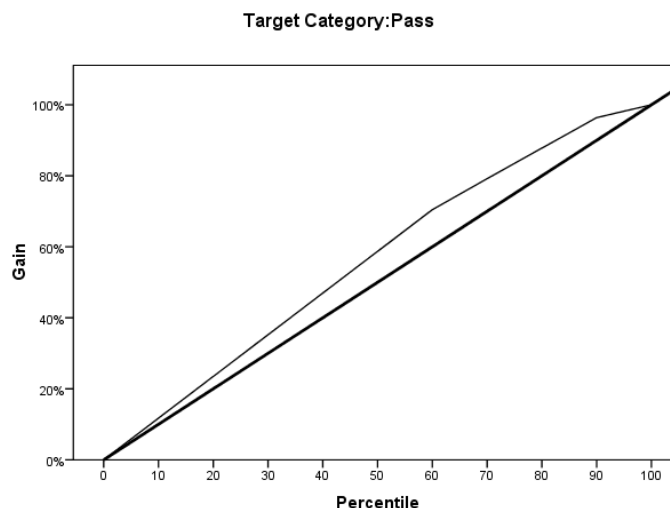
Figure 4: Gain chart for successful student (Pass category) - CHAID

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

Classification tree rules can be easy to explain and used with the newly enrolled student. They can be written in IF-THEN format. Rules for the CHAID classification tree are given for all three terminal nodes.

Node 1:

IF Ethnicity = "Maori" **OR** "Pacific Islanders" **THEN** Study outcome = "Fail" with probability 0.854

Node 3

IF Ethnicity = "Pakeha" **OR** "Others" **AND** Course programme = "Bachelor of Business" **THEN** Study outcome = "Pass" with probability 0.555

Node 4:

IF Ethnicity = "Maori" **OR** "Pacific Islanders" **AND** Course programme = "Bachelor of Applied Sciences" **THEN** Study outcome = "Fail" with probability 0.594

The CHAID classification tree only in Node 1 makes clear distinction between successful and unsuccessful students, because of the high probability (0.854) for students in this node of being unsuccessful. The results for this particular course, *Information Systems*, suggests that Maori and Pacific Islands students need additional learning support to increase their chance of successful completion of the course. For other two terminal nodes the cases are almost evenly split, i.e. there are almost equal chance for students in these nodes to pass or fail the course. In other words the model doesn't make an accurate prediction in these nodes.

CART

Figure 5 shows the CART classification tree for study outcome. It shows that only three variables were used to construct the tree: (1) ethnicity, (2) course programme and (3) course block.

The largest successful group (i.e. students who successfully completed the course) consists of 215 (47.5%) students (Node 5). The ethnic origin of students in this group is either Pakeha or other ethnic groups (excluding Maori and Pacific Islands students). Students in this group enrolled on the Bachelor of Business programme in either Semester 1 or Semester 2. The largest unsuccessful group (i.e. students who were unsuccessful) contains 138 students (30.5% of all participants) (Node 4). They belong to either Pakeha or other ethnic groups (excluding Maori and Pacific Islands students). The next largest group considered also as unsuccessful students, contains 41, i.e. 9.1% of all students, where 75.4% of them are unsuccessful (Node 1). They are described as Maori and Pacific Islands students.

Table 5: CART classification matrix

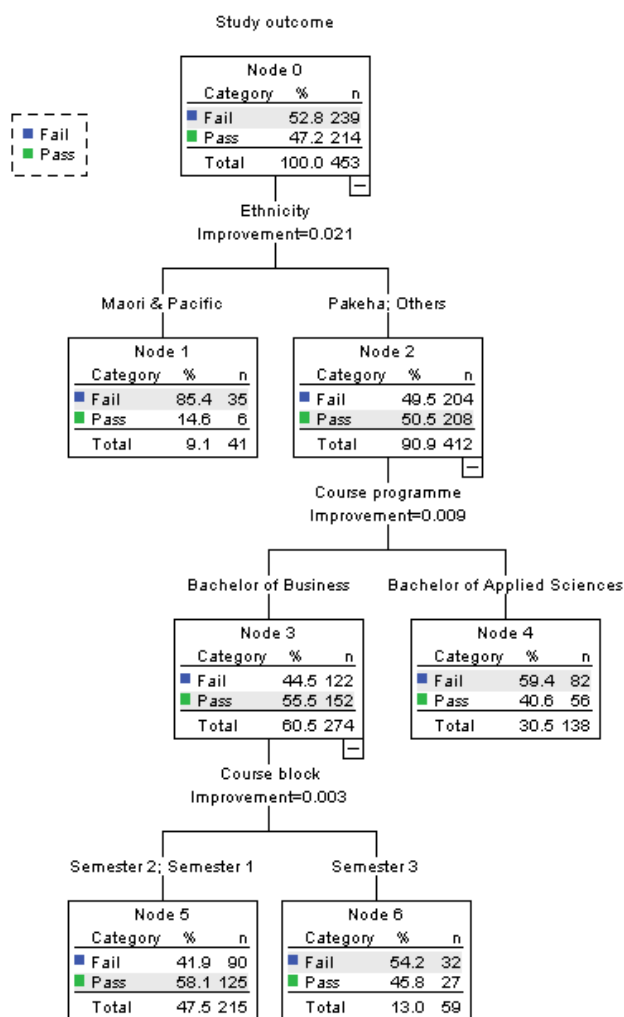|	Predicted		
Observed	Fail	Pass	Percent correct
Fail	149	90	62.3%
Pass	89	125	58.4%
Overall percentage	62.6%	58.1%	60.5%

Figure 5: CART classification tree

The cross-validation estimate of the risk is 0.446 indicates that the category predicted by the model (successful or unsuccessful student) is wrong for 44.6% of the cases. The CART classification matrix (Table 5) shows that model classify correctly 61% of students. This is slight increase in comparison to the CHAID model. The numbers of false positives (90) for the CART model decreases and therefore increasing the positive predictive value to 62.3%. In other words it will work better than the CHAID model at identifying an unsuccessful student. The price paid for increasing accuracy is reflected in decreasing sensitivity. The CART model will pick up 62.6% of all unsuccessful students (CHAID model 65.4%). At the same time the specificity will increase to 58.1% (CHAID model 55.5%).

The gains chart for the CART classification tree is presented in Figure 6. Gains charts for two models are almost the same, with slightly larger gains obtained for the CART model.
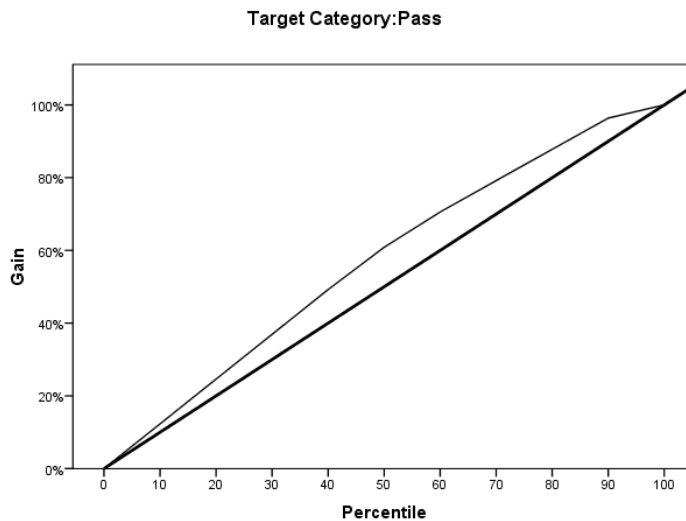
Target Category:Pass

Figure 6: Gain chart for successful student (Pass category) - CART

Rules for the CART classification tree are given for all four terminal nodes.

Node 1:

IF Ethnicity = "Maori" **OR** "Pacific Islanders" **THEN** Study outcome = "Fail" with probability 0.854

Node 4

IF Ethnicity = "Maori" **OR** "Pacific Islanders" **AND** Course programme = "Bachelor of Applied Sciences" **THEN** Study outcome = "Fail" with probability 0.594

Node 5:

IF Ethnicity = "Pakeha" **OR** "Other" **AND** Course programme = "Bachelor of Business" **AND** Course block = "Semester 1" **OR** "Semester 2" **THEN** Study outcome = "Pass" with probability 0.581

Node 6:

IF Ethnicity = "Pakeha" **OR** "Other" **AND** Course programme = "Bachelor of Business" **AND** Course block = "Semester 3" **THEN** Study outcome = "Fail" with probability 0.542

These rules could be used with a new data set to decide on the possible study outcome for a newly enrolled student.

The classification tree results for the study outcome on the *Information Systems* course suggest that the background information such as gender, age, disability and work status were not identified by the classification tree algorithms as factors of value in separating successful from unsuccessful students. The only significant demographic factors were ethnic origin and course related attributes such as course program and course block. However, these factors were not quite successful in identifying 'at-risk' students. These results are quite consistent with other publish 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 the demographic variables were used. Background characteristics could be significant initially, i.e. on their own, 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 when detecting study outcome or persistent student from dropout.

Concluding Remarks

This study examines the background information from enrolment data that impacts upon the study outcome of Information Systems students at the Open Polytechnic. Based on results from feature selection (Figure 2 and Table 3), the CHAID tree presented in Table 4 and Figures 3 and 4, the CART trees presented in Table 5 and Figures 5 and 6 it was found that the most important factors that help separate successful from unsuccessful students are ethnicity, course programme and course block. Demographic data such as gender and age though significantly related to the study outcome, according to the feature selection result, were not used in the classification trees. Unfortunately the classification accuracy from the classification trees was not very high. In the case of the CHAID tree the overall classification accuracy was 59.4% and in the case of the CART tree slightly higher at 60.5%. This would suggest that the background information (gender, age, ethnicity, disability, secondary school, work status, and early enrolment) gathered during the enrolment process, does not contain sufficient information for an accurately separation of successful and unsuccessful students.

Classifying students based on pre-enrollment information and the rules presented for each node 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, 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, this research is based on background information 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 classification trees. For example, including the assignment mark after the submission of the first course assignment would probably improve 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, we used a dichotomous variable for the study outcome with only two categories: pass and fail. However, splitting the fail category into those who stayed on the course but eventually failed the course and those who voluntary transfer or were withdrawn from the course would probably provide better profiling for each of the three categories of study outcomes. The only problem we might have with the three categories for study outcome is a low prediction accurate as a result of relatively small data set for the course. Thirdly, from a methodological point of view an alternative to a classification tree should be considered. The prime candidates to be used with this data set are logistic regression and neural networks.

Acknowledgement

I would like to thank John Green, Senior Lecturer in the School of Information and Social Sciences at the Open Polytechnic of New Zealand for valuable comments on the first draft of this paper. However, the authors should be held responsible for any remaining errors.

References

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.

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. *Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09)*. July 1-3, Cordoba, Spain, 41-50.

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

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

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.

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.

Reason, R. D. (2003). Student variables that predict retention: Recent research and new developments. *NASPA Journal, 40*(4), 172-191.

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.

Biography

Dr Zlatko J Kovačić is an Associate Professor in the School of Information and Social Sciences at the Open Polytechnic of New Zealand. Zlatko has a varied academic background and research interests, ranging from core interests relating to IT careers, learning and teaching, to eCommerce, eLearning, time series analysis, multivariate analysis and data mining. His current research is focused on modeling student retention, time series analysis of emerging financial markets, social and cultural aspects of information technology, and on cognitive processes in distance education using computers and communications technologies.