Systems Research and Behavioral Science
Syst. Res. **30**, 194–203 (2013)
Published online 30 August 2012 in Wiley Online Library
(wileyonlinelibrary.com) **DOI**: 10.1002/sres.2130

■ *Research Paper*

Predicting Students' Progression in Higher Education by Using the Random Forest Algorithm

Julie Hardman[1], Alberto Paucar-Caceres[1]* and Alan Fielding[2]

[1] *Manchester Metropolitan University, Business School, Manchester, UK*
[2] *Manchester Metropolitan University, School of Science and Environment, Manchester, UK*

This paper proposes the use of data available at Manchester Metropolitan University to assess the variables that can best predict student progression. We combine virtual learning environment (VLE) and management information systems student records datasets and apply the Random Forest (RF) algorithm to ascertain which variables can best predict students' progression. RF was deemed useful in this case because of the large amount of data available for analysis. The paper reports on the initial findings for data available in the period 2007–2008. Results seem to indicate that variables such as students' time of day usage, the last time students access the VLE and the number of document hits by staff are the best predictors of student progression. The paper contributes to VLE evaluation and highlights the usefulness of RF, a technique initially developed in the field of biology, in evaluating an educational and learning environment. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords evaluation; virtual learning environment; management information systems; student progression; Random Forest

INTRODUCTION

Student progression is an increasingly important issue for university institutions in the face of constrained funding for higher education in many countries around the world. In the UK, student retention varies from 81.6% to 98.8% depending on the university (National Audit Office, 2007).

It seems that there are many measures of undergraduate student progression (Kaighobadi and Allen, 2008), but the contention of this paper is that universities do not leverage the data that might be available to them within their own computer systems that give vital clues to student progression. Numerous sources of data exist in higher education institutions, from records of achievement to student finance and from staff personnel details to library loans. Each of these datasets is owned and used by different departments across the university campus and holds a wealth of information about every aspect of life

*Correspondence to: Alberto Paucar-Caceres, Manchester Metropolitan University, Business School, Manchester, UK.
E-mail: a.paucar@mmu.ac.uk

Received 15 November 2011
Accepted 5 July 2012

within the institution and beyond. These datasets have, in the past, resided in isolated silos, sometimes unavailable for any other use than that for which it was intended. However, in recent years, there has been a move towards interoperability with the advent of institution-wide systems, such as a managed learning environment. It can be argued that a university's business is very much about students progressing through their academic careers; therefore, it is of the most crucial importance to ascertain the factors affecting this progression and whether the available data can give insights into factors that can assist successful student progression.

The purpose of this paper is to explore the potential of using existing datasets to give possible indicators to predict successful student progression. The datasets focused upon in this paper are from two of the most ubiquitous systems within a UK higher education institution, namely the virtual learning environment (VLE), with particular emphasis on student usage data, and the universities management information systems (MIS), focusing on student records, with particular interest in data on student progression. Manchester Metropolitan University (MMU) embarked upon a project to roll out a VLE across all *seven* university campuses in 2005–2006. To date, approximately 20 000 of the 35 000 students across the institution use the VLE on at least one unit, with some campuses having a presence on all units studied by the students. The authors play a part in the team responsible for implementing, and more crucially evaluating, the VLE. The research project is ongoing and is in its fourth year of operation/implementation at MMU. One of the areas of particular importance is the evaluation of the VLE system. This paper represents one part of that evaluation focused on student progression. The research questions driving this paper are summarized as follows:

How can existing student datasets be leveraged within a university setting to provide indicators of student progression?
How can the data available be collated and analysed to find predictors of student progression?

Consequently, the aims of this paper are twofold: (1) in the context of VLE, we propose a framework that will combine two datasets in order to categorize relevant variables to student progression; and (2) by using a statistical tool [Random Forest (RF) algorithm], the paper reports on the initial analysis carried out to identify variables that best predict student progression at MMU.

The paper is organized as follows. The next section provides a literature review covering some definitions related to student progression and e-learning systems. The third section describes the methodology for the study; the fourth section provides some initial findings from datasets available at MMU. Finally, we discuss our initial findings and conclude by suggesting that progression factors are firstly more complex than at first glance and that all universities could leverage their existing data to gain insights on progression.

VIRTUAL SYSTEMS ENVIRONMENT IN HIGHER EDUCATION AND STUDENTS' PROGRESSION

For a majority of students in higher education in the UK, VLEs have become a standard part of their learning resources. VLE refers to the system that provides online interactions of various kinds between learners and tutors, including online learning (University of Brighton, 2003). This aspect of the system is often referred to in the literature under a whole host of titles including e-learning (Devedzic, 2003), online learning (Myers *et al*., 2004), computer-assisted learning (Laurillard, 1977) and computer-mediated learning (Alavi, 1994). In a recent Universities and Colleges Information Systems Association (UCISA) survey, 96% of institutions were found to have a VLE in use (UCISA, 2008).

Each institution within the UK will have its own database storing student details including personal information such as addresses, course information and the units the student is studying. It is a requirement for all institutions within the UK to report on their activities including detailed information about its students and the units they take. These information are critical as the institutions' funding is based upon returns made from these data.

Factors affecting student progression have long been the focus of research projects across higher education. Many of the published research projects have focused on a whole host of factors that affect the students' academic success and hence

DOI: 10.1002/sres

progression through their academic life and beyond (Kaighobadi and Allen, 2008).

There are some differences between what is understood as student progression and student retention. In the UK, progression and retention of students in higher education is high up the agenda of universities and advisory bodies. The National Audit Office produces a report every two years, monitors retention and noncontinuation, and makes recommendations as to how to improve the completion of students in higher education (National Audit Office, 2007). These are some of the definitions used by the National Audit Office (2007, pp. 53–54): *(a) Continuation: The proportion of the annual intake of new students who return to higher education in the subsequent year. (b) Completion: For the purposes of the performance indicators published by the Higher Education Statistics Agency, completion refers to the proportion of new students projected to obtain a degree at their original institution within 15 years*.

For this paper's purposes, the notion of *student progression* refers to the successful completion of an academic year. *Conversely not progressing* means the student would remain at the institution but be required to retake aspects of their studies and so not progress normally through their studies. Retention, or nonretention, on the other hand, involves the student leaving the institution completely regardless of their academic performance.

In a recent paper, Kaighobadi and Allen (2008) found research into factors for successful student progression clustered around intellectual and nonintellectual aspects. Intellectual factors include SAT scores, high school grades and other various grades achieved in an assortment of academic activities. Nonintellectual factors, on the other hand, focuses on demographic (e.g. gender, age and race), behavioral (time spent on exam preparation and outside work activities) and personality (e.g. motivation and confidence). None of the studies highlighted by Kaighobadi and Allen (2008) looked at the relationship between the students' use of e-learning systems and progression. Nor have any studies of this nature been found in the search for literature surrounding this area. This paper attempts to fill this gap. Figure 1 presents the general conceptual map underpinning our approach.

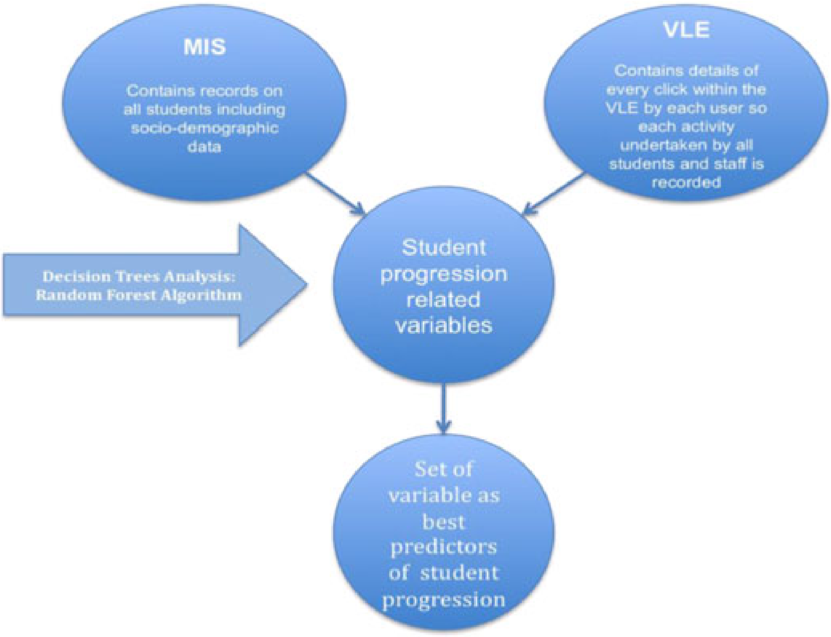
MIS
Contains records on all students including socio-demographic data
VLE
Contains details of every click within the VLE by each user so each activity undertaken by all students and staff is recorded
Decision Trees Analysis: Random Forest Algorithm
Student progression related variables
Set of variable as best predictors of student progression

Figure 1 Conceptual diagram of the study: combining virtual learning environment (VLE) and management information systems (MIS) datasets to identify relevant variables to indicate successful student progression

DOI: 10.1002/sres

PREDICTING STUDENT PROGRESSION USING EXISTING DATASETS: RANDOM FOREST ANALYSIS

The conceptual framework of the study is illustrated in Figure 1. The general aim is to coherently make use of the datasets available, VLE and MIS, and extract from them the variables we believe are relevant to student progression. The VLE at MMU contains details of every click within the VLE by each user, because each activity undertaken by all students and staff is recorded. The VLE dataset is a rich repository of student interaction with the learning system, presenting information on e-learning usage patterns, including how often it is used and when and what activities are undertaken. The data held within the VLE database can be utilized to highlight which usage variables best predict successful student progression. On the other hand, MIS contains records on all students including socio-demographic data and, importantly for this study, data on progression. Although open access was not granted to the MIS database, the data relevant to this study were made available. For VLE data, open access was granted to a reporting database at the end of each academic year.

From each dataset, variables that were believed to be relevant to student progression were extracted, resulting in a single dataset of 'student progression-related variables', as is shown in Figure 1. A number of variables were highlighted as being important to this study from both the VLE and MIS databases. The final variables, presented in Table 1 and used in the analysis, were formed as an aggregation of well over a million of rows of data held in the VLE and were brought together with the MIS database for the academic year 2007–2008; this is a consequence of the

Table 1 Student progression related variables

Variables used for analysis	Description	Dataset
Student progression	Students' progression as a Boolean variable—yes or no	SRS
Number of learning contexts student registered on	The number of different areas on WebCT Vista that a student is registered on; a learning context can be a unit/course	VLE
Number of active learning contexts student registered on	The number of learning contexts a student has been active on, that is. those units/courses with WebCT usage	VLE
Total number of staff registered on the students' learning contexts	The total number of staff registered on all their learning contexts	VLE
First used VLE	The total number of days from the start of term when the student first accessed VLE	VLE
Last used VLE	Total number of days from the start of term when the student last accessed VLE	VLE
Percentage usage between 9 AM and 9 PM	The percentage of individual student's usage between 9 AM and 9 PM	VLE
Percentage usage between 9 PM and 9 AM	The percentage of individual student's usage between 9 PM and 9 AM	VLE
Number of distinct student sessions	Total number of times the student accessed the system	VLE
Number of student document hits	Total activity for the student in accessing documents and resources	VLE
Number of student chat hits	Total activity for the student in accessing the chat/forum functionality	VLE
Number of student assessment hits	Total activity for the student in accessing the assessment areas	VLE
Total staff document hits	Total number of staff document hits for the staff registered on the students' learning contexts	VLE
Total staff chat hits	Total number of staff chat hits for the staff registered on the students' learning contexts	VLE

SRS, student record system; VLE, virtual learning environment; MIS, management information systems.

diversity of activities, comments, queries, and so on that 35 000 students generate when they access the VLE system. The myriad of interactions in the VLE activity log (ranging from students logging in to see a past exam, to converse or try to receive advice from tutors) were amalgamated into the 14 variables shown in Table 1; the first variable, student progression (yes/no), was known at the end of the academic year and came from the student record system dataset. The other 13 variables represent our attempt to amalgamate the myriad of activities, coming mainly from the usage of around 35 000 students when logging into the VLE system. The dataset was created using a bespoke SQL-stored procedure to interrogate the database and extract the requisite variables from the vast amount of data available. Because of the sheer magnitude of the data created for an academic year (2007–2008), we needed a simple algorithm able to cope with the number of items generated by the combination of 35 000 students and 13 variables for each student. After careful consideration, we decided to use a classification/decision tree algorithm, which is being increasingly widely used in academic disciplines, called *Random Forest* (RF).

Random Forest analysis (Breiman, 1999; Fielding, 2006) is a classification method based on the notion of decision trees; it has been widely used across many disciplines for treating large sets of data. The results and output of the RF analysis, in the form of a set of variables ranked according to its power to best predict student progression, are discussed in the last part of the paper. Because of space limitations, we cannot elaborate on the method in detail, but interested readers are referred to Breiman (1999, 2001a, 2001b); Breiman and Cutler (2004a, 2004b) and Fielding (2006). Because of the size of the data we handle here, and our desire to develop a predictive model, this was an appropriate method. The algorithm was developed at Berkeley, University of California. According to its creator, Breiman (1999:1), 'Random Forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all the trees in the forest'. Essentially, the classification method uses decision trees as its basis. The algorithm generates many random trees to form a forest. Trees are random because they use a random subset of the data and a random sample of the predictor variables to generate predictions for cases that were withheld, the so-called out-of-bag sample. Consequently, no two trees in a forest are identical. By growing each tree to its largest extent possible, with none of the pruning used in standard decision trees, helps to keep the bias low. According to Brieman and Cutler (2004b; quoted in Fielding, 2006), it has a number of important features:

- They have an accuracy that equals or exceeds many current classifiers, and they cannot overfit the data.
- They are efficient (fast) on large databases and can handle thousands of predictors without the need for variable selection routines.
- They estimate the importance of each predictor.
- They generate an unbiased estimate of the generalization error.
- They have robust algorithms to deal with missing data and for balancing the error if class proportions are markedly different.
- Generated forests can be saved for future use on other data.

Estimating variable importance is one of the RF's most important properties. Variable importance is estimated by determining the percentage increase in prediction error, arising from the exclusion of a predictor variable. Thus, predictors with larger values for their importance statistics contribute most to the correct prediction of the class of the out-of-bag cases. In our analyses, the predicted class is a binary variable that records a student's progression (yes or no). The resultant model can be expressed as student progression being predicted by the various ways in which students use/engage the VLE facilities available. Following RF bagging by which successive trees do not depend on earlier trees (each is independently constructed using a bootstrap sample of the dataset), we assume independence between the students' progression and the various events related to students' usage of VLE facilities. As it is indicated by the model depicted in Figure 1, we look to rank the importance of the variables in Table 1 as best predictors of student progression.

DOI: 10.1002/sres

FINDINGS

Table 2 shows the outcome of the analysis and ranks the variables by importance in predicting successful progression. The variables are highlighted in order of importance.

As can be seen from Table 2, the variables with the greatest importance to student progression are the following:

- percentage of students' usage between 9 AM and 9 PM;
- number of days from the start of term when the student last accessed VLE; and
- total number of staff document hits for the staff registered on the student learning contexts.

Next, we review the particular effect that these variables have, after accounting for the joint effect of the other predictors, in influencing student progression. A series of partial dependence plots were produced.

Daytime usage of the VLE appears to be a strong probability indicator for student success as shown in Figure 2. Daytime usage is classified as being between the hours of 9 AM and 9 PM. A daytime-to-nighttime usage ratio of approx 90% : 10% appears to give the maximum probability of success.

The probability indicator that ranked second is the time when the student accessed the VLE last. This is counted in the number of days since the start of term. Figure 3 shows a surge in probability of success based on the last time the VLE was used. This peaks towards the end of the spring term. For reference, 250 days from the start of the recording process is around the beginning of May.

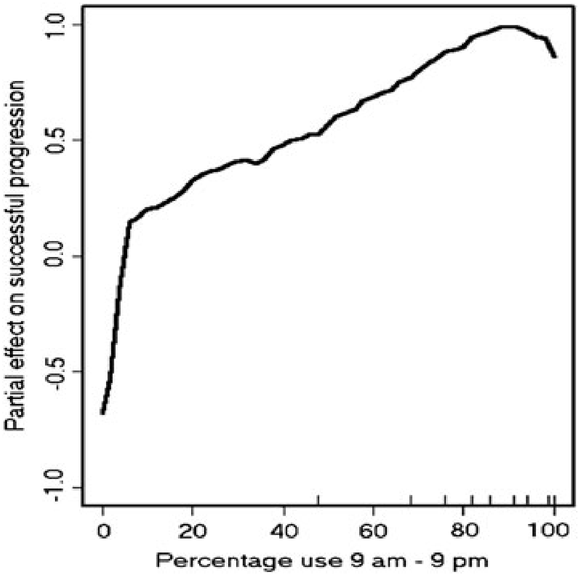

Figure 2 Partial dependence plot showing the effect of time of day of use on student progression

An interesting observation concerns the variable staff document hits, which consists of the total number of hits, for all staff registered on all the learning contexts (units, course, etc.) for each

Table 2 Importance of variable to predicting successful progression

Order of importance	Variable used for analysis	Progression indicator in order of importance
1	Percentage of student usage between 9 AM and 9 PM	906
2	Number of days from the start of term when the student last accessed VLE	768
3	Total number of staff document hits for the staff registered on the students' learning contexts	609
4	Number of student document hits	409
5	Total number of staff chat hits for the staff registered on the students' learning contexts	398
6	Number of distinct student sessions	394
7	Number of days from the start of term when the student first accessed VLE	386
8	Number of student assessment hits	317
9	Total number of staff registered on the students' learning contexts	301
10	Number of student chat hits	284
11	Number of learning contexts student registered on	233
12	Number of active learning contexts student registered on	227

DOI: 10.1002/sres

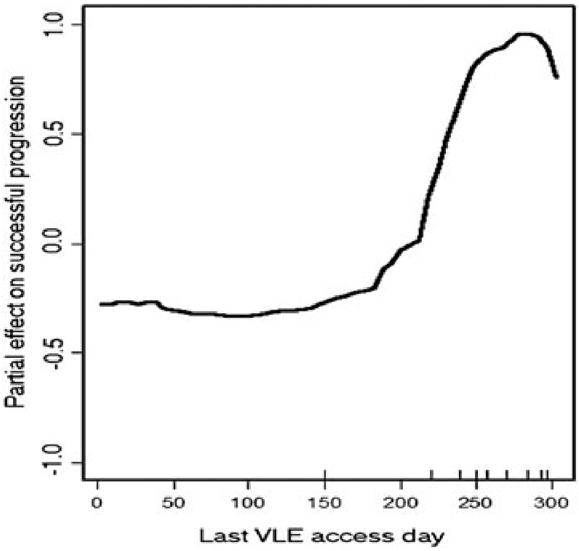

Figure 3 Partial dependence plot showing the effect of last access of virtual learning environment (VLE) on student progression

student, and resource provision-related activities such as uploading documents (e.g. lecture slides/ notes) and creating external links. Therefore, if a team of staff teaches a module, the hits are the total for the teaching team. As can be seen in Figure 4, the probability of students' success reaches a peak at a relatively low number of hits.

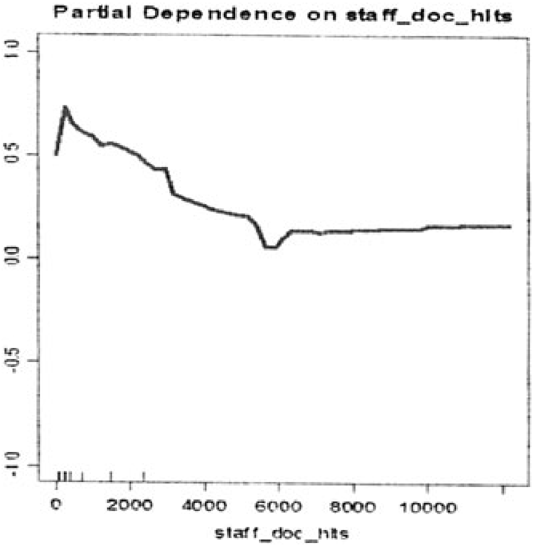

Figure 4 Partial dependence plot showing the effect of total staff hits to documents across all the students' learning contexts on student progression

DISCUSSION OF INITIAL FINDINGS

There appears to be some strong indicators to successful student progression in the usage patterns of the VLE by both staff and students. Each of these groups is an end user of the VLE but with very different agenda. For staff, the VLE is a teaching resource, whereas for students, it is a learning resource.

It is unlikely that any one factor highlighted is responsible for a student's progression independently; however, the purpose of this study is to gain some understanding of the usage profiles of staff and students and its possible relationship to student success. In understanding this, it may help the institution to highlight those students most at risk to not progress.

It is interesting to note some of the indicators that are most likely to predict student progression.

Student Usage

The way a student uses the VLE appears to have the greatest impact on the probability of their successful progression. The top two variables in Table 2 are related to student usage.

Looking at the variable 'usage by time of day', 9 AM to 9 PM users of the VLE are the most likely predicted to progress compared with those who principally use the system 9 PM to 9 AM. As was shown in the graph (Figure 2), there is a steady rise in the probability of success. Those that are least successful appear to use the VLE between 9 AM and 9 PM approximately 5% of the time rising to those who are most successful, using it 9 AM to 9 PM approximately 90% of the time. This study has yet to explore the reasons for this, and this could form the basis of further research. Many questions are raised with these findings, such as the following: What are the characteristics of the students who work the majority of their time at night? Are there similarities in the subject area for those students who prefer nighttime study? Are there issues in the students' personal lives that force them to access the VLE at night? Does accessing the VLE at night also mean that those are the hours the student studies? Many assumptions could be made, but only

DOI: 10.1002/sres

through further research will the reasons why students mainly access the VLE between 9 PM and 9 AM emerge.

The last time during the academic year a student accesses the VLE appears to be an important indicator to successful progression. The closer to the exam period the student last accessed the VLE, the more likely the student was to successfully progress. The exam period at MMU ran from 14 April to 16 May in the summer of 2008, which ties in with the rise on the graph. The assumption, when looking at these data, is that this variable is associated with utilizing the VLE for exam revision. The expectation would be that those students who use all available resources to assist them in their assessments do better than those who do not in the run up to an examination. However, this may not be the case, and this assumption could be flawed. At this stage of the research project, it is not known what resources were available to the students who did not access the VLE in the period immediately before their examinations; in other words, we do not know what resources they may have garnered offline. Questions raised may be related to students' VLE usage, such as the following: Did the students use the VLE to aid their examination revision? If not, why not? The questions could equally have no relationship to students' VLE usage and relate to wider issues such as the following: Were resources available within the VLE for their units? Is the timing of the examination the factor here with students having greater success in the later exams than the earlier ones? Only further research will help the understanding of the role their use of the VLE around the time of the examinations plays in their eventual progression.

Staff Usage

The results for the probability of student progression, influenced by the number of hits to document resources by all the staff involved in the modules, are interesting. It would appear to show that more staff activity for document-related activities has a detrimental effect on student progression. This is a counterintuitive finding and warrants further investigation. It is unknown at this stage whether the level of activity equates to the level of content. The point at which staff activity maximizes student progression appears to be low, which raises a number of questions. What relationship does activity (number of hits) have to content? Assuming there is a direct relationship between these issues, when activity is low, does staff target resources around the assessment, hence leading to better student performance? If there are many resources, is it a case of students being unable to critically appraise the content to target for themselves those that are most useful? Does it mean that too many resources cause the students to miss important resources placed there to help them? Only further research will assist in answering these questions, and it is intended that further analysis of this phenomenon will be undertaken.

CONCLUSIONS AND FURTHER RESEARCH

This research project is a preliminary foray into the data that already exist in higher education institutions in the UK and could be used for better understanding of student progression. It has provided a first look into whether usage of the VLE by academic staff and students can be used as a predictor to successful student progression. Because of the complexity of the subject under study, the results have provided more questions than answers. Each of the variables used have been pursued in isolation, and it is understood that multiple factors may affect students' successful progression. However, this study is a preliminary look at the individual variables on staff and students' usage of the VLE that leads to the highest probability for student progression. The initial findings show that there are some variables that are important predictors to successful student progression. These include students' time of day usage; the last time students access the VLE within an academic year; and the number of document hits by staff. This suggests that usage of the VLE by students and staff plays an important part in the final outcome for a student.

The ubiquitous nature of the datasets used in this study across UK higher education institutions means that 'mash ups' of existing data such as these are possible for the majority of institutions. This first look at the data has shown that the findings are not clear-cut, and further investigation into each finding is required. It has, however, proved to be an interesting study with some noteworthy findings and has now provided a springboard for further research. We contend that this paper makes two contributions—first, we defined variables for student progression in datasets that already exist on institutional systems and can therefore be leveraged by most institutions; and second, we illustrated the utility of RF analysis in assisting the analysis of those variables.

We are aware that some questions arising from the study are unanswered. Each variable that was tested raised issues that require further explanation of the results gained. General questions that could lead to further research include the following: Is the same phenomenon observed longitudinally? If not, what are the differences? Are there other student VLE usage patterns that have not currently been explored but could be derived from the database?

There are also some questions related to 'student usage by time of day' such as the following: What are the characteristics of the students who work the majority of their time at night? Are there similarities in the subject area for those students who prefer nighttime study? Are there issues in the students' personal lives that force them to study at night?

Questions related to 'student usage by last time the VLE was used' seem also relevant of a further study, and these include the following: Did the students not access the VLE at this critical time? Were resources available on the VLE for their units? Is the timing of the examination the factor here with students having greater success in the later exams than the earlier ones?

And finally, questions related to 'staff document activity' include the following: What relationship does activity have to content? Assuming there is a direct relationship between these issues, when activity is low, does staff target resources around the assessment, hence leading to better student performance? If there are many resources, is it a case of students being unable to critically appraise the content to target for themselves those that are most useful? Does it mean that many resources cause the students to miss important resources placed there to help them?

A possible starting point for further research will be to repeat the analysis with a new dataset; the intention being to collect longitudinal data over academic years. Currently, data have been analysed for academic year 2007–2008 and form the basis of this study. Data for 2008–2009 have now been made available, and so further analysis will be taking place in the coming months on that dataset. This will allow a comparison to confirm whether the same observations exist from one dataset to another.

Student progression is a critical issue to all higher education institutions, and we urge our fellow researchers to consider innovative use of existing datasets to provide a multifaceted view of student progression, as opposed to single measures. Insights gained in this area may ultimately highlight those students at risk from nonprogression. Only by exploring the datasets already held in the majority of universities will these interesting, valuable insights be gained.

REFERENCES

Alavi M. 1994. Computer-mediated collaborative learning: an empirical evaluation. *MIS Quarterly* **18**(2): 159–174.

Breiman L. 1999. Random Forests—random features. Statistics Department, University of California, Berkeley, Technical Report 567, September 1999.

Breiman L. 2001a. Random forests. *Machine Learning* **45**: 5–32.

Breiman L. 2001b. Statistical modelling: the two cultures. *Statistical Science* **16**: 199–215.

Breiman L, Cutler A. 2004a. Interface Workshop, April 2004, http://www.stat.berkeley.edu/~breiman/RandomForests/interface04.pdf, visited 1st February 2006.

Breiman L, Cutler A. 2004b. Random Forests. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm, visited 1st February 2006.

Devedzic V. 2003. Think ahead: evaluation and standardisation issues for e-learning applications. *International Journal of Continuing Engineering and Lifelong Learning* **13**(5/6): 556–566.

DOI: 10.1002/sres

Fielding AH. 2006. *Cluster and Classification Techniques in the BioSciences*. Cambridge University Press: Cambridge.

Kaighobadi M, Allen MT. 2008. Investigating academic success factors for undergraduate business students. *Decision Sciences Journal of Innovative Education* **6**(2): 427 – 436.

Laurillard D. 1977. Evaluation of student learning in CAL. *Computers in Education* **2**: 259–265.

Myers C, Bennett D, Brown G, Henderson T. 2004. Emerging online learning environment and student learning: an analysis of faculty perceptions. *Educational Technology & Society* **7**(1): 78–86.

National Audit Office (NAO). 2007. Staying the course: The retention of students in higher education. Report by the controller and auditor general. The Stationary Office, London, 26 July 2007. URL: http://www.nao.org.uk/idoc.ashx?docId=f2e92c15-d7cb-4d88-b5e4-03fb8419a0d2&version=−1 Last accessed: 24th February 2010.

UCISA. 2008. 2008 Survey of Technology Enhanced Learning for higher education in the UK. URL: http://www.ucisa.ac.uk/publications/~/media/groups/tlig/vle_surveys/TEL%20survey%202008%20pdf.ashx Last accessed: 25th February 2010.

University of Brighton. 2003. Managed Learning Environment Activity in Further and Higher Education in the UK. URL: http://www.jisc.ac.uk/index.cfm? name = project_mle_activity Accessed: 1st February 2006.

DOI: 10.1002/sres