**UNIVERSITY OF SOUTHAMPTON**

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING

Electronics and Computer science

# Linked Data Technologies to Support Higher Education Challenges: Student Retention, Progression and Completion

by

**Farhana Sarker**

Thesis for the degree of Doctor of Philosophy

April 2014

**UNIVERSITY OF SOUTHAMPTON**

# ABSTRACT

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING

Electronics and Computer Science

Doctor of Philosophy

## LINKED DATA TECHNOLOGIES TO SUPPORT HIGHER EDUCATION CHALLENGES: STUDENT RETENTION, PROGRESSION AND COMPLETION

by Farhana Sarker

Around the world, higher education institutions are facing a growing number of challenges. In recent decades, considerable interest has emerged on identifying those challenges and proposing efficient ways to address them. This thesis reviews a wide range of literature on higher education challenges and identifies related intuitional data, data repositories and external open data sources to address these challenges. It subsequently explores whether certain higher education challenges and in particular student retention, progression and completion can be better addressed using data from various data sources and the recent development of technologies such as, data analytics and linked data. Traditionally, research in this area is survey-based and survey-based studies have some drawbacks such as, low participation rate and the high cost associated with it. This research sought to overcome these problems. To this end, two experiments were conducted. The first experiment examined the sufficiency of linked data and external open data sources to develop blended prediction models to predict at-risk students in their first year of study. The result based on 149 undergraduate students' data, established that prediction models based on institutional repositories and external open data perform better than survey-based one. The second experiment examined the capabilities of institutional repositories and external open data sources in predicting students' first year marks and established that models using institutional repositories and external open data sources can perform better than models based on only institutional repositories. In order to examine the capabilities of linked data, external open data and data analytics, a data integration and analytics environment was deployed. The four key contributions of this thesis are: (1) it presents a comprehensive list of higher education challenges and required data and data repositories to address these challenges; (2) it demonstrates how external open data sources can be used to accurately predict students at-risk and students' first year marks; (3) it shows how including external open data sources in prediction models can increase the overall model accuracy and (4) it establishes the strengths and weaknesses of linked data to support in employing data analytics for predictive models in student retention, progression and completion.

# Table of Contents

# List of Tables

# List of Figures

# Declaration of Authorship

I, Farhana Sarker, declare that this thesis and the work presented in it are my own, and has been generated by me as the result of my own original research.

**LINKED DATA TECHNOLOGIES TO SUPPORT HIGHER EDUCATION CHALLENGES: STUDENT RETENTION, PROGRESSION AND COMPLETION**

I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- none of this work has previously been submitted for a degree or any other qualification at this University or any other institution;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- this thesis is based entirely on work done by myself and none of it was done jointly with others;
- none of this work has been published before submission, except the following:

  I. **Sarker, F.**, Davis, H. and Tiropanis, T. 2010a. A Review of Higher Education Challenges and Data Infrastructure Responses, *In the proceedings of 3rd International Conference of Education Research and Innovation (ICERI2010)*, Madrid, Spain, 1473-1483.

ii. **Sarker, F.**, Davis, H. and Tiropanis, T. 2010b. The Role of Institutional Repositories in addressing Higher Education Challenges. *In the proceedings of Second International Workshop on Semantic Web Applications in Higher Education,* University of Southampton, UK.

iii. **Sarker, F.**, Tiropanis, T. and Davis, H. 2012. A Linked Data Infrastructure to Support Student Retention, Progression and Completion, *Multidisciplinary Postgraduate Research Showcase 2012*, University of Southampton, UK, 2012.

iv. **Sarker, F.**, Tiropanis, T. and Davis, H. 2013a. Exploring Student Predictive Model that Relies on Institutional Databases and Open Data Instead of Traditional Questionnaires. *In the proceedings of 3rd International Workshop on Learning and Education with the Web of Data (LILE2013), WWW2013 Conference.* Rio de Janeiro, Brazil, 413-418.

v. **Sarker, F.**, Tiropanis, T. and Davis, H. 2013b. Students' Performance Prediction by Using Institutional Internal and External Open Data Sources. *In the proceedings of 5th International Conference on Computer Supported Education (CSEDU2013).* Aachen, Germany, 8.

*vi.* **Sarker, F.,** Tiropanis, T. and Davis, H. 2014. Linked Data, Data Mining and External Open Data to Predict at-risk Students. *In the proceedings of IEEE-2$^{nd}$ International Conference on Control, Decision and Information Technologies, CoDIT'14. Metz, France, Nov 03-05, 2014 (invited paper).*

Signed:

Date:

# Acknowledgements

First of all, I am especially grateful to my supervisors, Dr. Thanassis Tiropanis for his constant support, guidance and encouragement to carry out my PhD research study and Prof. Hugh Davis for his guidance and suggestion during the period of my study.

Further, I would like to thank Dr. Su White and Dr Nicholas Gibbins for their valuable guidance provided as my internal examiner during my exams and later on.

Also, thanks to all my friends and other PhD students in the WAIS Lab, University of Southampton to whom I could always go when I needed any support and also for making me pleasant to work in this lab.

Sincere thanks to Helen Smith from classification and harmonization unit (CHU), office for national statistics (ONS) for clarifying the method they used in deriving socio economic class using in their dataset. I also like to extend my gratitude to Jenny Skeet, statistical analyst, Higher Education Statistics Agency for her coordination in clarifying the classification method of participation neighborhoods.

I would also like to thank Christopher Gutteridge for some discussions on RDF, SPARQL and University of Southampton's datasets.

Finally, and very importantly, I would like to thank my family for supporting me throughout the process despite the difficulties it brings to a young family of four.

# Abbreviations

| | |
|---|---|
| AA | Academic Analytics |
| ACT | American College Testing |
| BIS | Department for Business, Innovation and Skills |
| CATPCA | Categorical Principal Component Analysis |
| CETIS | Centre for Educational Technology and Interoperability Standards |
| DDD | Data Driven Decision |
| DT | Decision Tree |
| HE | Higher Education |
| HEA | Higher Education Academy |
| HEFCE | Higher Education Funding Council for England |
| HEI | Higher Education Institutions |
| HESA | Higher Education Statistics Agency |
| HTTP | Hyper Text Transfer Protocol |
| IR | Institutional Repositories |
| JISC | Joint Information System Council |
| LA | Learning Analytics |
| LD | Linked Data |
| LOD | Linked Open Data |
| LR | Logistic Regression |
| NAO | National Audit Office |
| NN | Neural Networks |
| NSS | National Student Survey |
| OD | Open Data |
| ONS | Office for National Statistics |
| QAA | Quality Assurance Agency |

| | |
|---|---|
| RAE | Research Assessment Exercise |
| REF | Research Excellence Framework (REF) |
| RDF | Resource Description Framework |
| SAT | Standardized Admission Test |
| SE | Student Engagement |
| SEC | Socio Economic Class |
| SOC | Standard Occupational Classification |
| SPARQL | SPARQL Protocol and RDF Query Language |
| SRPC | Student Retention, Progression and Completion |
| SVM | Support Vector Machine |
| UK | The United Kingdom of Great Britain and Northern Ireland |
| URI | Universal Resource Identifier |
| USA | The United States of America |
| VLE | Virtual Learning Environments |

# Chapter 1: Introduction

## 1.1 Overview

There are a variety of factors, such as wide and diversified student population, rapid development of information technology, increased societal expectations, economic and pressure from government that cause a increasing number of challenges in higher education institutions (HEI) around the world. In recent years, extensive work has focused on identifying those challenges, identifying opportunities and ways to address them.

Advancement of new technologies is changing higher education institutions, as they become just one source among many for ideas, knowledge and innovation. Since 2007, widespread interest in analytics has been increasing in higher education sector (van Harmelen and Workman, 2012). According to EDUCAUSE, analytics is the use of data, statistical analysis, and explanatory and predictive models to gain insights and act on complex issue (Bichsel, 2012). In the higher education institutions, complex issues are student retention, progression and completion, student recruitment, student satisfaction, finance and budgeting and many more. These complex issues have started to be investigated using analytics. Analytics in the education domain is providing increased opportunities to learning and teaching, and offers more convenient evidence based decision-making tool. In the education domain, generally two types of analytics are employed: learning analytics and academic analytics. Learning analytics are the application of analytic techniques for educational data, including data about learner and teacher activities, to identify patterns of behaviour and provide actionable information to improve learning and learning-related activities (van Harmelen and Workman, 2012). On the other hand, academic analytics are the analytic activities that are not strictly learning analytics, but it helps educational institutions to fulfil their mission in many areas of higher education, such as student recruitment, finance and budgeting (van Harmelen and Workman, 2012). Alongside the wide opportunities in using analytics in higher education, there exists challenges to achieve the success in growing use of analytics in higher education; one of such challenges is data. In van Harmelen and Workman's (2012) study found that certain aspects of data, such as data quality, data ownership, data access and data standardization can act as a barrier to analytics.

The types of data used in analytics are changing. Higher education institutions are collecting larger volume of data about its members such as, students and teachers, its facilities and curricula than ever before[1] (Campbell and Oblinger, 2007; Long and Siemens, 2011; Bichsel, 2012). Technologies are playing a central role in increasing this large amount of data that are continuously generated by people (Reinsel *et al.*, 2007). A study conducted by IDC in 2008 on the amount of existing digital data found that the rapidly expanding "Digital Universe" was expected to grow to 1.2 million petabytes (PB), or 1.2 zettabytes (ZB) in 2010 and to reach 35 ZB by 2020 (Reinsel *et al.*, 2007). Repositories are advances to efficiently store and access this large volume of data. It is argued that institutional repositories (IR) are a very powerful idea that can serve as an engine of change for institutions of higher education (McCord, 2003). If properly developed, they advance a surprising number of goals, and address an impressive range of challenges. Apart from the institutional data, in the United Kingdom (UK) a number of external bodies routinely publish educational open data in the web, such as the Higher Education Statistics Agency (HESA[2]), the Higher Education Funding Council for England (HEFCE[3]), the Office for National Statistics (ONS[4]), and Unistats[5]. The Open Data Institute (ODI[6]) defines open data as the information that is freely available for anyone to use for any purpose. Open data ensures the data interoperability. Open data need to have a licence stating that the data is open data. Without a licence, the data can't be reused by anyone[7].

[1] http://blogs.cetis.ac.uk/cetisli/2011/12/14/big-data-and-analytics-in-education-and-learning/

[2] http://www.hesa.ac.uk/

[3] http://www.hefce.ac.uk/

[4] http://www.ons.gov.uk/ons/index.html

[5] http://unistats.direct.gov.uk/

[6] http://www.theodi.org/guide/what-open-data

[7] http://opendefinition.org/

The Centre for Educational Technology and Interoperability Standards (CETIS) "Analytics for the Whole Institution; Balancing Strategy and Tactics" paper Kay and van Harmelen (2012) states that analytics can provide the best where data from multiple locations can be joined together based on commonly agreed coding frames for key elements and data collection increases over time. At the same time, Arnold (2010) identifies several barriers in combining data into a common location such as different technology standards, lack of unique identifiers, and organizational restrictions of ownership and use of data. Linked data (LD) technologies are considered to be well suited for data integration while data is in different locations as linked data provides more expressivity of data and follows a unique structure of data. In their online tutorial "How to Publish Linked Data on the Web[8]", Chris Bizer *et al.* define linked data as a style of publishing and interlinking structured data on the web. Evidence from the literature (Tiropanis *et al.*, 2009a; 2009b; Tiropanis *et al.*, 2009c; Tiropanis *et al.*, 2009d), linked data technologies are promising in addressing many higher education challenges as it is able to join data from disparate data sources. In his report "Linked Data Horizon Scan" Paul Miller (2010) also points out many opportunities of linked data in higher education and in their article " How Open Data, data literacy and Linked Data will revolutionise higher education" McAuley *et al.* (2011) stated the opportunities of linked data and open data in higher education institutions.

In a recent EDUCAUSE survey, evidence shows that currently most institutional data are primarily used for reporting and credential purposes rather than to address strategic problems (Bichsel, 2012). The CETIS analytics series demonstrates that there are increasing opportunities for the higher education sector to use analytics to produce innovative and meaningful ways to evidence performance and success of their institutions. Analytics can be applied to most strategic area of higher education, such as student retention, progression and completion, finance and resource allocation (Bichsel, 2012; van Harmelen and Workman, 2012).

In this thesis, we reviewed a broad range of literature and present a comprehensive list of current higher education challenges those are significantly impacting the higher education institutions to maintain their on-

[8] http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/

going progress, such as widening participation, student retention, progression and completion, higher education funding, quality of learning and teaching, quality of research, plagiarism, assessment and feedback. Also, we present the related data to address these higher education challenges in a structured way. Evidence from the literature (Tinto, 2006-2007), student retention, progression and completion is a widely researched area in the area of higher education for many years. It has become one of the major issues to be addressed by many UK higher education institutions (HEI) due to its influence in positioning institutions in league tables, and the complexities of the factors involved (Tinto, 1975; Walker, 1999; Thomas, 2002; Lowe and Cook, 2003; Yorke and Longden, 2004). Higher education institutions are taking a number of steps to increase student retention, progression and completion in their institutions. Early prediction of students' performance, which can be measure through course marks is one of the ways that help higher education institutions to undertake timely and pro-active measures for the poor performing students (Kotsiantis *et al.*, 2004; Kovacic, 2010). Once recognized these poor performing students can be then targeted with academic and administrative support to increase their chance of retention and completion of their course.

In this thesis, we aim to explore whether the emergent environment of linked open data and higher education repositories can address student retention, progression and completion adequately or even better than traditional approaches that rely on the resource-intensive completion of questionnaires.

## 1.2 Motivation

This doctoral research was initially motivated by the work conducted by the researchers through the SemTech project (Tiropanis *et al.*, 2009d). In this project, the researchers investigated the potential of semantic web technologies in supporting higher education challenges. After reviewing the literature on student retention, progression and completion, we found that, traditionally, research in this area is survey-based where researchers use questionnaires to collect student data to analyse and develop student predictive models to identify at-risk students. After identifying these at-risk students, they can arrange additional support for them to retain and succeed in their study. Though the survey-based model has been successfully used from many years, the main problem with the survey-based study is low participation rate, which may often compromise the precision of the output. Moreover, survey-based research may be too burdensome to sustain, as individual institutions may not have the capacity to construct and administer a similar instrument to study their unique retention situation. Even if an institution is capable of fielding a one-time retention survey, repeated administrations of such surveys can be costly. Thus, it is pivotal for the higher education institutions to source efficient means of developing student predictive models in order to develop and/or adjust support programs timely and accurately without having to compromise the precision of the model accuracy. Therefore, innovative techniques are in demand to develop predictive models to support student retention, progression and completion. One efficient way to develop a predictive model is make use of commonly available institutional internal databases and external open datasets. As Schwartz *et al.* (2010) note, data informed decision making helps higher education institutions know whether they are achieving their missions in different areas of HE such as increasing student success in degree completion. Institutions routinely collect a broad array of information on their students' backgrounds and academic progress. Also in the UK, a number of external bodies routinely publish some open datasets, which could be used to develop student predictive models in the place of questionnaire-based predictive models that have been used to-date.

Though a large amount of institutional internal and external open data is available, significant institutional challenges obstruct the implementation of analytics efforts since data are frequently maintained in different locations (Arnold, 2010). Linked data are well suited for data integration and interoperability from different data sources. Linked data seems promising in addressing many higher education challenges (Tiropanis *et al.*, 2009a; 2009b; Tiropanis *et al.*, 2009c; Tiropanis *et al.*, 2009d; Miller, 2010). Also McAuley *et al.* (2011) established the opportunities of linked data and open data in higher education institutions in their article "How Open Data, data literacy and Linked Data will revolutionise higher education".

In this context, this doctoral work aims to explore whether student retention, progression and completion can be better supported by integrating related data from disparate data sources (internal or external) and analysing new sets of linked data to more accurately predict at-risk students and students' marks in their first year of study.

Specifically, the **research questions** that motivate this doctoral research include:

- What institutional data/repositories can be used to efficiently address student retention, progression and completion? Are the currently available data sufficient to address this challenge?
- Are linked data technologies well suited to address this challenge? What are the advantages of using linked data technologies in this respect? Can we show how student retention, progression and completion can be efficiently addressed by aggregating information using linked data technologies from internal or even external data sources?
- Can we provide an infrastructure to efficiently monitor any potential data patterns that indicate stay/drop in student retention, progression and completion? What would be the challenges to provide such an infrastructure?

The following **hypotheses** are constructed from the above research questions:

- It is possible to provide accurate/improved student prediction models by combining institutional internal databases and external open data sources.
- Linked data can provide sufficient support for building student prediction model when combining institutional internal/external data sources.
- Institutional internal/external data sources can be used to compensate the lack of questionnaire data in building student prediction model.
- It is possible to predict students' mark using institutional internal and external data sources.

## 1.3 Approach

There are three stages employed to investigate the above mentioned research questions. In the first stage, we identified a number of higher education challenges and related data and data repositories (internal/external data sources) to address those higher education challenges from the literature. In reviewing literature, we have also documented the growing opportunities of linked data technologies and data analytics in higher education sector.  Based on this literature review, we constructed three research questions for this dissertation.

In the second stage, to respond to the research questions we deployed a linked-data based experiment architecture, which is able to connect to multiple data repositories (internal/external), perform SPARQL query over them and combine the query results into a single dataset. This single dataset is the required final dataset to build the predictive models.

In the final stage, we conducted two experiments to respond to the research questions. The first experiment seeks to examine: (a) whether institutional internal and external open data sources can be used in developing student predictive model to identify at-risk students without having to rely on traditional questionnaires and (b) whether including external open data sources in the predictive model increase the precision of the overall model performance.

The undertaking of the first experiment requires establishing a baseline based upon existing student predictive models that use traditional questionnaire. This is then compared these findings to our constructive student predictive model, which uses available institutional internal and external open data sources that does not rely on the traditional questionnaires.

To conduct this experiment, data were collected from two disparate sources. The first source was an online questionnaire comprising of 49 questions and divided into six parts (see *Appendix A* for details), to collect Institutional internal dataset items and traditional questionnaire items. Organising this online survey/questionnaire required an ethics approval from the University of Southampton. We applied for the ethics approval to the University of Southampton's ethics committee and obtained the approval from the

University's ethics committee (see *Appendix B* for reviewed documents). A total 149 students' data were collected. The second source of data, external open data, was Unistats website which publishes National Student Survey (NSS) feedback on students' satisfaction on their courses. Logistic regression models were developed to identify at-risk students in their study using 149 undergraduate students' data and NSS data. We applied logistic regression as most retention studies adopt this approach (Pascarella *et al.*, 1983; Langbein and Snider, 1999; Light, 2000; Herzog, 2005; Miller and Herreid, 2008; Singell and Waddell, 2010). Also, logistic regression is an established method in retention studies for it handles both categorical and continuous predictor variables, which do not have to exhibit linearity and homogeneity of variance vis-a`-vis the outcome variable (Hosmer and Lemeshow, 2000; Peng *et al.*, 2002). The results of this experiment demonstrates that a predictive model using institutional internal databases and external open data sources can provide better performance compared to the survey-based model or traditional questionnaire based model and the model solely based upon institutional internal databases. This finding supports the hypothesis that student predictive models can be developed using institutional internal databases and external open data sources without having to rely on questionnaire data, which have been traditionally used in retention studies for many years. Moreover, this finding supports the hypothesis that predictive model including external data sources can perform as same as or can out-perform the traditional survey-based predictive models.

A second sought to examine (a) whether institutional internal and external open data sources can be used to predict students' first year mark and (b) whether including external open data sources in the predictive model can increase the precision of the overall model performance. In the second experiment, we developed four predictive models. The first model uses only students' background data from institutional internal datasets. The second model uses both institutional internal and external datasets. The findings of these two models then compared to find out the best performing model among them. The third and fourth models are developed by adding students' current academic performance (first semester mark) with the preceding two models to examine the effects of adding students' first semester marks to the model performance. We use decision tree to develop the models. It is found that decision tree is popular in predicting students' academic performance (Al-Radaideh *et al.*, 2006; Bharadwaj and Pal, 2011a; Yadav *et al.*, 2011; Yadev and Pal, 2012). A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. The results of the second experiment illustrates that including external open data sources can provide better prediction performance than compared to the model based on only institutional internal data sources.

## 1.4 Contributions

This research study contributes as follows:

**Contribution 1**: from the literature review on higher education challenges, this research identifies a number of challenges in the higher education. This investigation has been published (Sarker *et al.*, 2010a) in the proceedings of 3rd International Conference of Education, Research and Innovation (ICERI 2010). Moreover, it is noted that Laureate Education, Inc.[9] has been using this publication as a course material in their online program since November 3, 2011.

**Contribution 2**: from the literature review on the higher education challenges this research identifies the data and data repositories are required to address these higher education challenges. This work has been published (Sarker *et al.*, 2010b) in the proceedings of the second international workshop on Semantic Web Applications in Higher Education (SEMHE 2010).

**Contribution 3**: this research explores improved student progression prediction models using institutional internal and external linked open data sources without having to rely on traditional questionnaire those are essential in developing prediction model. This investigation has been published (Sarker *et al.*, 2013a) in the proceedings of 3rd International Workshop on Learning and Education with the Web of Data (LILE2013), WWW2013 Conference.

**Contribution 4**: this research explores improved students' mark prediction model using institutional internal and external linked open data sources. This investigation has been published (Sarker *et al.*, 2013b) in the proceedings of 5th International Conference on Computer Supported Education (CSEDU2013).

[9] http://www.laureate.net/

**Contribution 5**: this research presents the suitability of linked data technologies in supporting student retention, progression and completion.

**Contribution 6**: this research presents the sufficiency of external open data sources in developing student predictive model without having to rely on traditional questionnaires.

## 1.5 Organization of the thesis

The rest of this thesis is organized as follows:

**Chapter 2** engages with the literature to provide an overview of higher education challenges and it elicits 22 key challenges facing today's higher education institutions.

**Chapter 3** provides literature review on student retention, progression and completion. First, it discusses different definitions of student retention, progression and completion. Second, it discusses how higher education institutions monitor student retention, progression and completion. Third, it explores why it is important and what use HEI make of the collected data. Fourth, it presents related data and datasets. Fifth, it considers five well-known retention theories/models and discusses a number of studies based on Tinto's model, which is the most widely accepted model in the retention literature. Sixth, it discusses studies on students' mark/performance prediction. Finally, it points out some issues with the traditional survey-based prediction models.

**Chapter 4** presents the concepts and the rationale for linked data technologies in higher education, open data and the role of data analytics in higher education.

**Chapter 5** elicits the related institutional internal datasets and external open data sources to address the higher education challenges in a structured way with specifying the merits and demerits of sharing repositories/data sources.

**Chapter 6** explains the experimental design and methodologies with providing an ontology for student retention progression and completion. Also, it provides an overview of the two experiments that have been conducted to test the hypotheses. Moreover, this chapter presents the linked-data experimental environment that has been employed for integrating data from multiple data repositories (internal/external).

**Chapter 7** reports the first experiment in detail with it’s result, which explores a new way of developing student predictive models to identify at-risk students in their study by using institutional internal datasets and external open data sources without having to rely on traditional questionnaires.

**Chapter 8** presents the second experiment and it’s result, which predicts students’ academic performance in their first year of study based on institutional internal datasets and external open data sources.

**Chapter 9** discusses each hypothesis separately, stating the support for each claim as it stems from the experiment. Also, it states the limitations of the studies.

Finally, **Chapter 10** outlines the major findings of this PhD research study and specifies future research based on the study results.

This thesis also includes six appendices that complement the discussion of the thesis. *Appendix A* collects the questionnaires used to collect student data and *Appendix B* contains ethics review documents for the questionnaires. *Appendix C* presents a list of variables related to student retention, progression and completion. *Appendix D* includes the classes and properties of the student retention progression and completion (SRPC) ontology. *Appendix E* includes more detail about the variables used in experiment 1 and *Appendix F* presents the full derivation of students’ parents annual mean income and students’ socio economic class based on students’ parents’ occupation from office for national statistics (ONS) published open datasets.

# Chapter 2: Higher Education Challenges

Higher education is vital to the continued economic growth and prosperity of the nation, creating a skills base, which allows people to get on and into employment. Higher education includes teaching, research, exacting applied work, and social services activities of universities. The beginning of the 21st century has been a period of expansion in higher education in most of the countries in the world. Over the past decade, higher education attainment has increased by almost 10% across OECD countries (OECD, 2013). At the same time, in many countries preferably in the western higher education where the education was free before, governments plan to cut their contributions and introduce variable/increased tuition fee on higher education. These, along with a number of other changes, have affected almost every aspect of higher education provision and have served as a precursor to the current reforms of the higher education sector.

In the UK, the student population has grown remarkably during the last decade. According to HESA statistics, during the period 2002–03 to 2012–13, the total number of students at higher education institutions in the UK increased by almost 210,000, or 10%. However, the rate of change fluctuated within the period: 2009–10 saw the largest year-on-year increase, of 4.1%, and 2012-13 saw the largest year-on-year decrease, of 6.3% in the total student population (see figure 2.1, data source: HESA). This change in the student population can be coincided: the strongest increases in the population (in 2009–10) coincided with the beginning of the economic downturn and the significant decrease (in 2012-13) in the student population is due to the introduction of increased tuition fee up to £9000. The effect of governments funding cut and the introduction of variable tuition fee can also be seen in 2007-08 with a decreased of 2% in the student population while the government introduced tuition fee (up to £3000) in 2006-07. Therefore, it can be said that the decreased in the student population are more pronounced due to the plan of governments to cut their contributions to higher education and introducing tuition fee. However, a much wider range of factors is almost certainly at play, shaping year-to-year volatility in total numbers such as, the regulation on student number control and the change of immigration policy in the UK.

It is noticed that over the same period, the sector has experienced increasingly international outlook. The growth in total student numbers has largely come from a significant increase in the number of international students studying at UK universities. Moreover, the UK's share in the international student market grew from 10.8% in 2000 to 13.0% in 2011 (Universities UK, 2013). UK is the second preferred destination among international students after USA. In 2002–03 non-EU students made up just 8.5% of the total student population, which had risen to 12.8% by 2012-13 (figure 2.2, data source: HESA). International students bring huge benefits to the UK economy. It is noted that in 2011–12 international students spent about £10.2 billion on tuition fees and living expenses. Figure 2.3 (data source: HESA) shows the trends of funding body grants (for teaching and research) and tuition fee income from 2002-03 to 2012-13. From the figure, it can be observed that there is a decreasing trend of public funding over the time period while there is an increasing trend of students tuition fee income. The decreasing trend of public funding can also be observed elsewhere in the OECD.

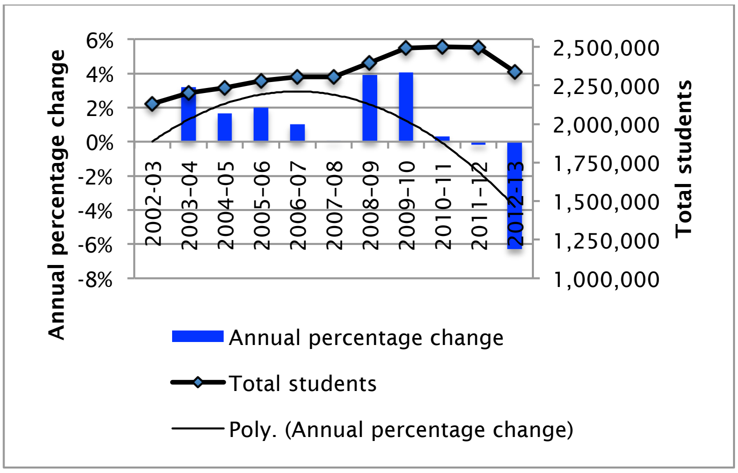


Figure 2.1. Total number of students with annual change, 2002-03 to 2012-13. (Source: HESA)

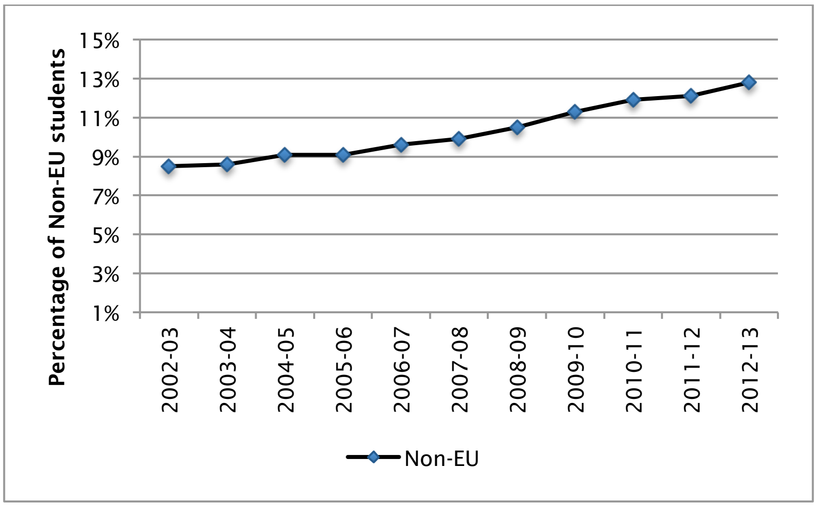


Figure 2.2. Proportion of Non-EU/International students, 2002-03 to 2012-13. (Source: HESA)

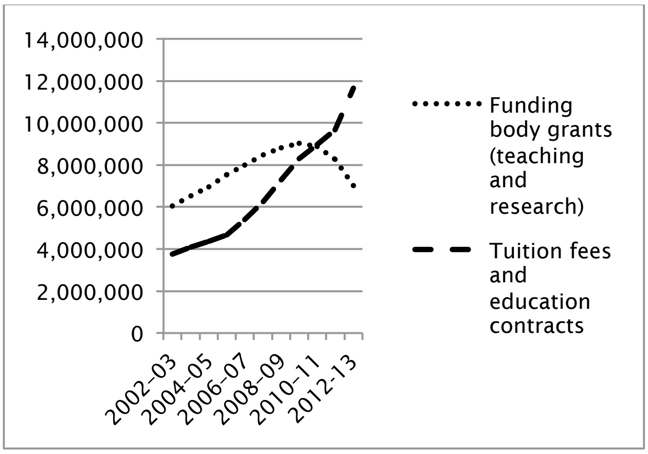


Figure 2.3. Trends of public funding and tuition fee income, 2002-03 to 2012-13. (Source: HESA)

Overall, it can be said that higher education has undergone significant change, from the introduction of variable and increased tuition fees to the expansion in the number of student population. Due to these along with other changes higher education is facing a number of challenges not only in the UK but also around the world. In this chapter, we review challenges faced higher education sector around the world while providing more emphasis in the UK higher education context and present a comprehensive list of those higher education challenges found in the literature.

## 2.1 Search Methods and Information Sources

To refine the online search regarding higher education challenges, the following keywords were used to search the Google scholar, ISI Web of Knowledge and University of Southampton's online library resources: higher education challenges, higher education issues, global problems in higher education and higher education challenges in UK. All articles' bibliographies were searched for additional resources. In addition, given priority to the news, reports and publications of different stakeholders of the higher education sector. In the context of higher education, stakeholders mean specific groups of actors that have a direct or indirect interest in higher education; such as government, students, quality assurance agency, accreditation bodies, academic staff and employers. A list of key stakeholders in the UK higher education are identified based on their activities and reliability in providing higher education related statistics and information. Their websites provide additional information about their organization, their published reports and publications; and other information of interest to the higher education community.

- **The Higher Education Funding Council for England (HEFCE):** HEFCE is the funding body in the UK which distributes public money for teaching and research to 130 universities and Higher Education colleges. It promotes high quality of education and research within the Higher Education Institutions (HEI). HEFCE is available at http://www.hefce.ac.uk.
- **The Department for Business, Innovation and Skills (BIS):** BIS is building a dynamic and competitive UK economy by creating the conditions for business success; promoting innovation, enterprise and science; and giving everyone the skills and opportunities to succeed. To achieve this, it fosters world-class universities and promote an open global economy. BIS is available at http://www.bis.gov.uk/.
- **The National Student Survey (NSS):** The aim of NSS is to gather feedback on the quality of students' courses in order to contribute to public accountability, as well as, to help inform the choices of future applicants to higher education. The NSS is the largest survey of its kind in the UK. Over the last 10 years it has helped over 2

million students to make their voices heard about the things that matter to them, and has been fundamental to driving change in universities and colleges. The result of this survey is published in unistats website, http://unistats.direct.gov.uk and the NSS website is http://www.thestudentsurvey.com/.

- **Quality Assurance Agency (QAA):** QAA is an independent body funded by subscriptions from universities and colleges and through contracts with the higher education funding bodies. They carry out external quality assurance by visiting universities and colleges to review how well they are fulfilling their responsibilities for academic standards and quality, identifying good practice and making recommendations for improvement. They also publish guidelines to help institutions develop effective systems to ensure students have high quality experiences. QAA is available at http://www.qaa.ac.uk/.
- **The Institution of Engineering and Technology (IET)**: IET is the second largest engineering institution in the world and is a professional body. The IET accredits degree courses worldwide in subjects relevant to electrical, electronic, manufacturing and information engineering. They are available at http://www.theiet.org/.
- **British Computer Society (BCS):** BCS is the chartered institute for IT. BCS is a professional body, a learned society, a nominated and awarding body. They bring together industry, academics, practitioners and government to share knowledge, promote new thinking, inform the design of new curricula, shape public policy and inform the public. They are available at http://www.bcs.org/.
- **Academic research paper:** Journal papers, conference papers and electronic books from different sources were reviewed. Some mentionable journals include: journal on Higher Education, the Institute of Electrical and Electronics Engineers (IEEE), electronic journal on e-Learning, journal of Education + Training, journal of Educational Policy, journal of Assessment and Evaluation in Higher Education, NACADA journal, journal of Digital Information, International journal on Semantic Web and Information Systems. Relevant workshop papers are also included such as: International Workshop on

Ontologies and Semantic Web for E-Learning, International Workshop on Semantic Web applications in Higher Education.

- **Higher Education Academy (HEA):** HEA is a national and independent organization, which is funded by mostly four UK funding bodies and by subscriptions and grants. HEA supports the higher education in order to enhance the quality and impact of learning and teaching. They are available at http://www.heacademy.ac.uk/.
- **Joint Information System Committee (JISC):** JISC inspires UK colleges and universities in the innovative use of digital technologies, helping to maintain the UK's position as a global leader in education. Their mission is to provide world-class leadership in the innovative use of Information and Communications Technology to support education, research and institutional effectiveness. They are available at http://www.jisc.ac.uk/.
- **The Council for Industry and HE (CIHE):** The CIHE is a strategic leadership network of businesses and higher education executives promoting a system of higher learning that leads to greater market competitiveness and social wellbeing. They are available at http://www.cihe-uk.com.

## 2.2 Higher Education Challenges

After reviewing the literature 22 higher education challenges are highlighted to represent the current challenges in higher education sector. The following subsections exemplify these challenges.

### 2.2.1 Higher Education Funding

Higher education institutions are in financial crisis. According to House of Commons Education and Skills Committee (2003) in the UK, funding per student fell 36% between 1989 and 1997, and the investment backlog in teaching and research facilities is estimated at £8 billion. In another report, the House of Commons (2013) stated that the funding council cut its funding by £449 million (about 6%) for financial year 2010-11 and total funding through the funding council in England was provisionally reduced by £680 (about 9.5%) in 2011-12 compared to 2010-11.

Many potential students with low socio economic conditions fail to enrol in higher education due to funding crisis. Moreover, increased student fees, substitutions of loans for grants, diminishing subsidies to student facilities and so on form a financial barrier to prospective students (Hirsch and Weber, 1999; BIS, 2009). In the UK, a decreased of 2% in the student population in 2007-08 is documented with the introduction of up to £3000 tuition fee in 2006-07 and a significant decreased of 6.3% in the student population in 2012-13 is due to the increased tuition fee up to £9000 in 2011-12 (Universities UK, 2013).

Funding crisis in higher education is in all over the world. According to American Council on Education *et al.* (2006) the federal government's Advisory Committee on Student Financial Assistance indicates that each year nearly 400,000 academically qualified students fail to pursue a postsecondary education because they cannot afford it. The authors in (Biggs and Tang, 2007) reported that 20 years ago public funding paid for virtually 100% of HE costs but today this is hardly the case, for example Australian students obtain only 30% of the university funding from the public purse. Therefore, it becomes a challenge to all HEIs to overcome this financial crisis.

### 2.2.2 Fair Admission and Widening Participation

Education is a fundamental principle to build a more social society as education is the best and most reliable route to scape out of poverty and disadvantage (Education and Skills Committee, 2003). The demands of higher education increased due to the increased opportunities of HE graduates in the labour market. It is recognised that graduates earn more than workers who do not have a higher education qualification (Universities UK, 2013). But the problem is access to higher education due to social origin, increased student fees, substitutions of loans for grants, diminishing subsidies to student facilities (Hirsch and Weber, 1999; Prime Minister's Strategy Unit, 2007). Too many lower income and minority students fail to enrol in higher education due to funding crisis (ACE *et al.*, 2006; BIS, 2009).

In many countries, the current pressure is to increase the number of students in the higher education. Since 1997, the UK government has increased funding to support the HE sector in widening participation (Prime Minister's Strategy Unit, 2007). Improving access and improving participation in higher education are also a crucial part of the Higher Education Funding Council for England's strategic mission. They believe that all those who have the potential to benefit from higher education should have the opportunity to do so. Department for Education and

Skills (2003) in the UK believe that admissions should always be on merit irrespective of socio-economic class, ethnic background or but rather based upon an applicant's achievements and skill potential. Moreover, to increase higher education participation, the House of Commons Education and Skills Committee (2003) and HEFCE (2010a) have advised that those who wants HE for the first time be given priority in accessing HE.

HEFCE also give emphasis on fair admission principles and suggest institutions to provide necessary information such as admission requirement, admission procedures and student completion rate in the course, so that prospective students can take their decisions accordingly (HEFCE, 2009a; The sub-committee for Teaching Quality and the Student Experience, 2009). The Joint Information Systems Committee (JISC[10]) is also working on improving fair admission and widening participation to support higher education institutions in the UK. They provide advice and guidance on how to ensure institution's provision to all kinds of students including disabled. They also work with university admissions service, University College Admission System (UCAS) to ensure the admission process fair with the help of technology. For example, they support applicants through the admission process, give feedback to unsuccessful applicants. QAA advised higher education institutions to provide personal support and guidance to students to choose their subject area for their higher study and also mentioned the necessity of collaborative activity with schools and colleges for increasing widening participation (QAA, 2008a; 2010).

### 2.2.3 Improving Student Retention, Progression and Completion

The types of student served by higher education institutions in the world have changed over time, moving from a small, selective, generally homogenous group of privileged individuals to a diverse spectrum of individuals numbering in the millions. According to Universities UK (2012a), the total number of students has grown by 28 per cent between 2000-01 and 2009-10, to roughly 2.5 million. As the student population has expanded and diversified, therefore student retention, progression and completion becomes an increasing complex issue in the higher education sector worldwide (Seidman, 2005). The National Audit Office (2007) reported around 28,000 full-time and 87,000 part-time undergraduates starting in 2004-05 did not continue to a second year of study. NAO also stated that thirty institutions experienced a fall of at least one percentage point in their continuation

[10] http://www.jisc.ac.uk

rates of first year students since 2001-02. Therefore, higher education institutions need to focus on improving student retention, progression and completion. The authors in the SemTech project (Tiropanis *et al.*, 2009d) and QAA (2008c) agree student retention, progression and completion as one of the most salient challenges in higher education institutions. Higher education institutions need to take it seriously to improve to remain best in the today's competitive world.

Higher education institutions are increasingly recognizing that student retention, progression and completion are cornerstones in solidifying public support, engagement and achieve a high standard of higher education institutions (Hanna, 2003). Higher education institutions require monitoring students' overall progress and based on this information they can take necessary steps for the students from the very beginning (QAA, 2008c; Tiropanis *et al.*, 2009d). From the very beginning, they also need to take into account why students dropout from a programme or from any specific modules. In the United States, it has been a documented issue in higher education since the late 1800's (Boston *et al.*, 2009). Many theoretical models have emerged to guide and assist in the understanding of the underlying reasons for early departure by students such as, Tinto's student integration theory and Astin's student involvement theory (Braxton, 2000). To this end, measuring the level of student engagement has become the latest focus of attention in higher education (Trowler, 2010) to improve student retention, progression and completion as Chen *et al.* (2008) reported that student engagement is positively linked to high grades, students' satisfaction, and persistence. The term 'student engagement' indicates students' attitude towards their institutions and their academic and non-academic activities. Kuh (2009) has defined student engagement as "*the time and effort students devote to activities that are empirically linked to desired outcomes of college and what institutions do to induce students to participate in these activities*." Moreover, students' engagement identifies the nature of the relationships between students, academics, university resources, and the studies. Identifying these relationships helps to provide insights into the potential of the relationship between students' engagement and their academic achievement. Therefore, to improve student academic achievements and successively to improve student retention, progression and completion, higher education institutions need to focus on increasing student engagement.

In addition, Education and Skills Committee (2003) recommended to establish strategic plan to improve those universities with unacceptably high dropout rates. HEFCE and QAA also mentioned the necessity of improving student retention, progression and completion and recommend Institutions to improve student retention, progression and completion by providing provision for disabled, part-time and mature students (HEFCE, 2009a; JISC, 2009b; HEFCE, 2010a; QAA, 2010). It is also recommended that higher education institutions require to provide quality education, student support with academic guidance and supervision to students at all level of degree completion.

### 2.2.4 Quality of Learning and Teaching

Higher education institutions will lose their potential students if they cannot assure high quality standards (Hirsch and Weber, 1999). Maintaining quality has the highest priority to any organization and it is mostly appropriate to the higher education institutions. Higher education institutions should care about the quality of learning and teaching because it is the only way to become recognized globally. To be a quality institution, higher education institutions are required to meet all the expectations (the key principles that are essential for the assurance of academic standards and quality), which are identified by the higher education community (QAA, 2012). For example, the expectation about learning and teaching which HEIs are required to meet is "*Higher education providers, working with their staff, students and other stakeholders, articulate and systematically review and enhance the provision of learning opportunities and teaching practices, so that every student is enabled to develop as an independent learner, study their chosen subject(s) in depth and enhance their capacity for analytical, critical and creative thinking*". The QAA carries out reviews to check whether higher education institutions are meeting the expectations[11]. The QAA assured that quality of learning and teaching is one of the top criteria to be a quality institution and also for allocation of HEFCE funding (QAA, 2008b; 2008d; HEFCE, 2010a).

To improve the quality of learning and teaching, higher education Institutions can enable access to learning and teaching material across institutions (Tiropanis *et al.*, 2009d). Therefore, students/learners can get more information about their subject area to learn as well as teachers can have more information to teach broadly in an area (Hirsch and Weber, 1999; Hanna, 2003; BIS, 2009). The HEFCE advised that higher education institutions need to take extra care to maintain the quality of learning and teaching to ensure the best possible student experience (HEFCE, 2009a; 2010a). Maintaining excellence in both teaching and learning is

[11] www.qaa.ac.uk/InstitutionReports/types-of-review

key to universities (BIS, 2009). In the USA, the Fund for the Improvement of Postsecondary Education (FIPSE) is introduced to improve students' learning (ACE *et al.*, 2006).

On the other hand, as student fees now involve a high proportion of funding, universities are expected to improve the quality of their teaching. Moreover, to attract international students, universities are in demand to provide higher standard of teaching (Biggs and Tang, 2007; BIS, 2009). Since 1997, the UK government has increased funding to improve the quality of learning and teaching in HE (Prime Minister's Strategy Unit, 2007). The English higher education funding council, HEFCE aims to ensure that all HE students benefit from a high-quality learning experience that fully meets their needs and the needs of society at large. In the UK, JISC is also working on improving the quality of learning and teaching with the help of information technology (JISC, 2009b). For example, they create a supportive learning environment by providing virtual learning environments (VLEs) and mobile technology at the HEIs in the UK.

### 2.2.5 Curriculum Design/Alignment

To ensure the quality of learning, all institutions need to emphasize redesigning of the curricula. It has been argued that higher education Institutions should listen carefully to the changing needs and expectations of the society. In the SemTech project (2009d), the authors reported curriculum design/alignment as one of the higher education challenges from learning and teaching perspective.

To act competitively in the global HE environment, the higher education institutions must offer programmes to students that will cover their needs and wishes, and they can also provide interdisciplinary programmes to meet the 21st century's HE demands (ACE *et al.*, 2006; Rae, 2007). Hirsch and Weber (1999) and Prime Minister's Strategy Unit (2007) suggested that universities should be more responsive when offering new study programme or course. In the UK, the department for Business Innovation and Skills (BIS) (2009) states that no students will lag behind through curriculum alignment and they assert that all student can compete equally in this globalization era. According to the National Student Survey (NSS)[12] report, students overall satisfaction on their courses was 81.1% in 2005, which reached to 85.0% in 2013 and 86% in 2014 (HEFCE, 2014). Every year this survey conducted and published so that institutions could benefit and make improvement based on this report. Also, quality of course is one of the top

[12] http://www.nss.ac.uk

considerations for allocating funding for higher education (HEFCE, 2010a; 2010b). It is recognised that in the present world the students are paying more so their demands have increased in courses and quality, and higher education institutions should respond to these demands (Education and Skills Committee, 2003; Biggs and Tang, 2007; HEFCE, 2009b). Higher education institutions need to reformat and reorganize courses, programmes, and structures to increasingly sophisticated and market-knowledgeable students (Hanna, 2003).

### 2.2.6 Student Employability

All over the world, employability remains high on higher education Institutions' agenda. People are seeking educational opportunities to survive in the world of work (West, 1999). As the financial burdens on students and graduates grow, HE graduates increasingly find gaining a degree as a necessary first step to starting their career and for this reason employability is a major and growing concern (Hirsch and Weber, 1999; Rae, 2007; DIUS, 2008). According to Biggs and Tang (2007), the new agenda for education is to provide education for market needs.

Employability has been defined as a set of skills, knowledge and personal attributes that make an individual more likely to secure and be successful in their chosen occupation (Hirsch and Weber, 1999; Biggs and Tang, 2007; Rae, 2007; BIS, 2009; HEFCE, 2009b). Employability is also defined by the learning outcome of a programme with parallel personal development such as work experience and extra-curricular activities.

Higher skills significantly influence life chances and earning potential. The choice of degree subjects and its relevance to the employment market is affected to some extent and higher education institutions require to respond to this by involving employers in course validation to ensure that academic standards meet employer requirements (Hanna, 2003; Rae, 2007; BIS, 2009). The QAA also advise HEI in the UK to involve employers in course design and student placement to enhance students' employability (QAA, 2008a; 2008e). Moreover, in the UK, the Higher Education Funding Council for England states that employers are responsible for offering work placement and practical experience for students. As such universities should become more flexible in providing employers' needs by including the employability skills in their curriculum. According to Bridges (2000), 21st century's curriculum should consider student employability seriously and include key skills such as team working, communication skills, presentation skills, information technology and critical thinking to promote student employability.

Therefore, higher education institutions are in demand to take necessary steps to address this issue immediately for the greater interest of students as well as for themselves.

### 2.2.7 Assessment and Feedback

Assessment is a key process in higher education because it provides learners with assessment of their mastery of the curriculum and illuminates their ability to progress[13]. To derive maximum learning benefit from assessment, students need to receive timely feedback in a manner that is supportive. In relation to the feedback process, both students and teachers often disappoint and frustrate. Students frequently criticize that feedback on assessment is unhelpful or unclear, and sometimes even worrying. In addition, students sometimes complain that the feedback is provided too late to use and do not provide any guidance to improve subsequent performance (Spiller, 2009). Moreover, students express more dissatisfaction with assessment and feedback than with any other aspect of their learning experience. According to National Student Survey (NSS) report, in 2005 less than 50% of full time  students were satisfied with assessment and feedback of HEIs and it reached to 72% in 2014 (HEFCE, 2014). On the other hand, faculties express frustration that students do not incorporate feedback advice into subsequent tasks to improve and are only concerned with the mark. Therefore, assessment and feedback process in HEIs demand serious consideration to improve. According to Thomas (2002) and Tiropanis *et al.* (2009d), now a days, assessment and feedback become a matter of concern for many higher education institutions as both are important parts of the learning process. HEFCE and QAA stated that to be a quality institution, higher education institution should have effective assessment and feedback mechanisms and also to deal with breaches of assessment regulations, and the resolution of appear against assessment decisions (Hart and Friesner, 2004; HEFCE, 2009a; The sub-committee for Teaching Quality and the Student Experience, 2009). JISC also stated that effective assessment has greater bearing on successful learning than almost any other factor.

JISC in the UK has been working in technology enhanced assessment for over a decade, promoting work on the technical and interoperability issues associated with on-screen testing, and the broader technical, pedagogical and institutional

[13] http://www.jisc.ac.uk/whatwedo/programmes/elearning/assessment/assessworkshops.aspx

considerations for the effective use of a wide range of technologies to support assessment and feedback.

### 2.2.8 Group Formation for Learning and Teaching

To improve learning and teaching group formation becomes an important consideration in today's HE environment. Currently students come from diverse communities and/or different countries to study. Moreover, in some cases, students complete a course or degree online regardless of time and place (virtual university). To have efficient learning and teaching, teachers often like to put students into groups to work together for any projects, to participate in different discussion forums, or even to make batches of students in order to study their performance on a certain task (Bridges, 2000; QAA, 2008a; Tiropanis *et al.*, 2009d). This group formation can be based on different criteria such as students coming from different cultures, different gender so that students can easily communicate to each other and improve their learning by efficiently sharing their knowledge.

### 2.2.9 Critical Thinking and Argumentation

In recent years, critical thinking has been recognized as an important aim of higher education institutions to improve students' learning (Bridges, 2000; Tiropanis *et al.*, 2009d). Critical thinking employs not only logic but also broad intellectual criteria such as clarity, credibility, accuracy, precision, relevance, depth, breadth, significance and so on. The process of critical thinking involves the careful acquisition and interpretation of information and use of it to reach a well-justified conclusion. Critical thinking is important, because it enables one to analyse, evaluate, explain, and restructure thinking. The various skills that are collectively termed 'critical thinking' are regarded as an important component of the so-called 'transferable skills' accrued during higher education (JISC, 2010). To build students perfectly for this competitive and demanding world HEIs need to give more emphasis on supporting their students in critical thinking.

### 2.2.10 Construction of Personal and Group Knowledge

In the todays collaborative learning environment, where the speed of knowledge creation is very high and demanding, higher education institutions also realising the importance of more personal and group knowledge creation (Bridges, 2000). Higher education institutions can focus on improving the quality of learning and teaching by more efficient personalized knowledge construction allowing access to the knowledge capitals of higher education institutions, as well as more efficient contextualized group knowledge construction (Hirsch and Weber, 1999; Hanna, 2003; Tiropanis *et al.*, 2009d). Hence, it becomes one of the aims of today's higher education institutions.

### 2.2.11 Integration of Knowledge Capital and Cross-curricular Initiatives

To support better learning and teaching activities, integration of HE knowledge capital like research output, learning and teaching materials and the alike is essential (Tiropanis *et al.*, 2009d). Also cross-curricular activity in learning and teaching, and in research is essential to improve the standard of the higher education institutions. According to BIS (2009) and Tiropanis *et al.* (2009d), cross-curricular activities in emerging areas by matching teachers to new programme and module definitely enhance the quality of learning and teaching in higher education institutions (BIS, 2009). Therefore, it becomes one of the most important target of today's demanding and diverse HE (Bridges, 2000).

### 2.2.12 Developing New Generation of Staff

The best-organized institution is worth nothing if it does not have a qualified teaching staff; an unqualified staff means poor teaching and unimaginative research (Hirsch and Weber, 1999; DIUS, 2008).

In order to successfully teach the curriculum including employability skills, universities need to develop the new capacities among their traditional teaching staff and new approaches to their teaching (Bridges, 2000). Higher education institutions need to develop faculty and staff dedicated to engaging a diversity of learners with more complex learning needs. Higher education institutions can offer different types of training for their staff so that they can be up to date with current HE environment and undertake professional development when necessary (Hanna, 2003; Biggs and Tang, 2007). In HEFCE's annual survey, more than 60% of institutions reported difficulties in recruiting lecturers. This is compounded by the fact that the average student:staff ratio across the sector increased from 9:1 in 1980 to 13:1 by 1990 and a further increased to 17:1 by 1999 (23:1 if funding for research which is included in the average unit of funding is excluded) (Greenway and Haynes, 2003).

New generation of faculty should consider professional development as a lifelong process. This means that they need to be up to date with the changing landscape of HE pedagogy. As per House of Commons Education and Skills Committee (2003), Centres of Excellence in Teaching will be established to reward good teaching at departmental level and to promote best practice. The National Teaching Fellowships Scheme will be increased in size to offer substantial rewards to twice as many outstanding teachers as at present. HEFCE (2009a) mentioned the necessity of training for faculty development as qualified staff can only provide quality of teaching. The QAA encouraged new generation of faculty to engage in research as part of their development (QAA, 2008d; 2010). The QAA also advise to engage in collaborative activities for faculty development of higher education institutions (QAA, 2008a). In the USA, the department of education developed a new programme "Preparing Tomorrow's Teachers to Use Technology" to develop new generation of faculty (West, 1999). In the UK, JISC also help in developing staff in UK higher education by providing resources and a range of activities including workshops, training to universities and their staff (JISC, 2009a).

### 2.2.13 Quality of Research

World-class research plays a key role in economic growth through creating new businesses, improving the performance of existing businesses, delivering highly skilled people to the labour market, and attracting investment from global businesses. To be the best worldwide in research, higher education institutions need to strengthen their research capacity. In order to achieve this challenge, higher education Institutions need to develop multidisciplinary centres with diverse and complementary expertise and build collaboration between universities and industries (BIS, 2009; HEFCE, 2009b).

In the UK, maintaining the quality in research is taken seriously and the government has increased funding for improving the quality of research (Prime Minister’s Strategy Unit, 2007). Maintaining quality requires a greater focus on world-class researchers and greater recognition of the potential benefits of research development in key area (BIS, 2009). The Higher Education Funding Council for England in the UK aims to develop and sustain a dynamic and internationally competitive research sector that makes a major contribution to economic prosperity, national wellbeing and the expansion and dissemination of knowledge. Also, distribution of HEFCE research funding across the HEIs focused on the best research, which was evaluated by Research Assessment Exercise

(RAE[14]), a peer review exercise (Education and Skills Committee, 2003; HEFCE, 2010a). Since 2008, Research Excellence Framework (REF[15]) a new system for assessing research has been introduced in the UK HEIs, which replaces RAE. REF results will be used by the funding bodies to allocate research funding to the HEIs from 2015-16. HEFCE invested more in the very best research institutions. According to Universities UK (2014), institutions in the fifth quintile (the upper 20% of the funding distribution) have received about 75% of quality related (QR) funding from funding councils in 2013–14.

### 2.2.14 Competing and Collaborating Globally in Research

There is global competition to attract and retain top talented students, researchers and lecturers (Prime Minister's Strategy Unit, 2007). Institutions need to compete at a world-class level in teaching and research. Higher education institutions need to maintain higher standard of research so that they can be recognized internationally and can compete with other HEI by means of higher quality and higher standard of research (Hirsch and Weber, 1999; BIS, 2009). Maximizing the research capacity HEI can make top quality relationships with other HE system elsewhere in the world (Education and Skills Committee, 2003; DIUS, 2008; HEFCE, 2009b). Moreover, higher education institutions are finding that international and local collaboration with other higher education Institutions, industry, communities and government is necessary to exploit the opportunities offered by globalization (Prime Minister's Strategy Unit, 2007). HEFCE gives priority in collaborative research for their funding distributions in HEI and also encourage universities to engage with business and communities for collaborative research (HEFCE, 2010a; 2010b).

HEI are expending more in their research to remain excellent and compete globally. On average across OECD countries, HEI spend 31% of all expenditure per student (OECD, 2013). Moreover, according to OECD (2013), the share of research and development expenditure as a percentage of GDP ranges from 0.05% in Brazil to 0.94% in Sweden. Surprisingly, it is noted that the HEI in the top two countries, USA and UK spend less in research and development, 0.31% and 0.46% of GDP respectively, while the average percentage of GDP across OECD countries is 0.45. Despite of the minimum expenditure, the UK research is top in the worldwide. The

[14] http://www.rae.ac.uk

[15] http://www.ref.ac.uk/

World Economic Forum evaluation (WEF) ranks the UK consistently in the top 5 countries for university-industry collaboration in research and development[16].

Moreover, it becomes a key concern in evaluating quality of HEI nowadays. In the UK, JISC works on how technology can efficiently support collaborative research (JISC, 2008; 2009b) in the HEI.

HEI are expending more in their research to remain excellent and compete globally. On average across OECD countries, HEI spend 31% of all expenditure per student (OECD, 2013). Moreover, according to OECD (2013), the share of research and development expenditure as a percentage of GDP ranges from 0.05% in Brazil to 0.94% in Sweden. Surprisingly, it is noted that the HEI in the top two countries, USA and UK spend less in research and development, 0.31% and 0.46% of GDP respectively, while the average percentage of GDP across OECD countries is 0.45.

### 2.2.15 Addressing Plagiarism

Recently, concerns have increased in the higher education system with regards to the incidences of plagiarism (the passing of someone else's work as though it was one's own work). Before 1990's occurrences of plagiarism were comparatively rare but the recent massification of higher education observable as a worldwide phenomenon, has raised concerns in the academic community that plagiarism may now be a serious and endemic problem (Hart and Friesner, 2004).

The SemTech project (Tiropanis *et al.*, 2009d) reported addressing of plagiarism as one of the vital issues in higher education. Carroll and Appleton (2001) believe that inaction in tackling the growing worries about and possible instances of plagiarism and collusion will threaten the integrity and reliability of higher education awards in the UK. HEFCE recommended that higher education institutions should have mechanisms to identify and deal with any academic misconduct such as plagiarism (HEFCE, 2009a; The sub-committee for Teaching Quality and the Student Experience, 2009).

In the UK, plagiarism is now considered sufficiently serious for academics to consult. Joint Information Systems Committee Plagiarism Advisory Service (JISCPAS[17]) established by JISC, promotes good practice in this area and provides guidance in all aspects of plagiarism prevention.

[16] WEF (2013) 12.04 University-industry collaboration in R&D

[17] http://www.jiscpas.ac.uk

### 2.2.16 Adopting Emerging Technology

Today's world is driven by technology for its communications, its economy and increasingly its day-to-day organization. The rapid development of information technology has made available a plethora of new tools for higher education (Fox, 1998; Hanna, 2003). New technology offers learning opportunities anywhere to anyone at anytime (Fox, 1998). Further, the response of higher education institutions to this new technology is uncharacteristically rapid. The lack of investment in technology based learning in higher education may prove to be a significant barrier to the ability of universities to compete in new or changing markets (Hanna, 2003; DIUS, 2008).

Technologies like Internet and its associated technologies can increase the capacity of an educator to quickly, easily and more palpably to aid students to make connections to content, context, and community resulting in more powerful learning experience (West, 1999). QAA's review process seriously takes into account the availability of learning and teaching technologies in the HEI. The accrediting agency for teacher preparation programs in the United States, NCATE is directly addressing the need for new faculties to be competent in the use of technology in their own teaching. This takes place by beefing up its standards for the year 2000 to take performance-based approach and will emphasize the use of technology aids (West, 1999). Thus, the need for the flexibility and contextual learning provided by technological tools is increasing. Higher education institutions should meet the challenge of technologies (Hirsch and Weber, 1999; Bridges, 2000; JISC, 2007; BIS, 2009). JISC is working to explore, test and acquire an understanding of a variety of technologies and how they might be used in HE (Anderson *et al.*, 2001; JISC, 2009b; 2009a).

### 2.2.17 Accreditation of HEI and Programme

One of the principal means of providing accountability for higher education institutions and their programmes is through accreditation. This is the most critical part of quality assurance in HE (ACE *et al.*, 2006; BIS, 2009). The Institution of Engineering and Technology (IET[18]) and British Computer Society (BCS[19]) stated that accreditation of degree programmes demonstrate institutions' commitment to developing and maintaining standards through. In 2008, Dr Andy Gravell, Director

[18] http://www.theiet.org
[19] http://www.bcs.org

of Undergraduate Studies, University of Southampton[20] stated: *"Professional accreditation and being able to successfully satisfy the standards of the accrediting bodies are extremely important for students. When they go out into the world of work, they can be assured that their degrees meet the highest professional standards."*

Accreditation is defined as a strong, meaningful assurance of academic quality (ACE *et al.*, 2006; Eaton, 2009). According to IET, accreditation can have several positive outcomes such as: assist to attract the best students, meet the needs of industry, benchmark programmes against other institutions both in the UK and internationally; and provide students with a good foundation for professional registration. Bridges (2000) adds accreditation affects institutions' ability to attract funding bodies or to attract interest from the business and private sectors. In the UK, HEFCE funding is available to higher education institution only if QAA and RAE/REF qualify the institution (The sub-committee for Teaching Quality and the Student Experience, 2009; QAA, 2010). Also Eaton (2009) stated that, in the USA, federal student aid funds are available to students only if the higher education institutions or programme they are attending is accredited by a recognized accrediting organization. Therefore, in order to attract students and funding accreditation becomes paramount to the higher education institutions. West (1999) and Tiropanis *et al.* (2009d) also specified accreditation as one of the major challenges in higher education institutions. All accreditors make students' learning outcomes a central component in the accreditation reviews (Hanna, 2003; ACE *et al.*, 2006; BIS, 2009). To efficiently accredit higher education institutions and programmes by professional bodies institutions can make related information accessible to the accreditation bodies. As institutions' information lies scattered across departments, so institutions can integrate that information and then make it accessible for efficient accreditation (Tiropanis *et al.*, 2009d).

### 2.2.18 Contribution to Economy

Institutions are seriously challenged to secure or even increase their revenues (Hirsch and Weber, 1999; DIUS, 2008). Higher education institutions pivotal in generating and preserving, disseminating and transforming knowledge for social and economic benefits (Prime Minister's Strategy Unit, 2007; BIS, 2009). It is vital that HEI use their knowledge capital to contribute to economic growth, both through the commercial application of the knowledge they generate and through

[20] http://www.ecs.soton.ac.uk

preparing people for the world of modern work (Education and Skills Committee, 2003; BIS, 2009).

HEFCE reported building new partnerships with business and industry provides an important channel for generating the financial resources (HEFCE, 2010a; 2010b). Also mentioned that over £2 billion a year in income is generated for HEI through knowledge and expertise collaboration with business and the wider community where income from knowledge exchange has been increasing at a rate of 12 per cent per year (HEFCE, 2010b). Further, HEFCE has allocated £27 million from the Economic Challenge Investment Fund to encourage HEI working with vulnerable people during the recession to contribute in the social economy. Higher education as an export industry already contributes around £8.3 billion to the UK economy, and this is expected to rise to around £17 billion by 2025 (Universities UK, 2012b).

According to BIS (2009), higher education Institutions need to give priority to the programmes that meet the need for high level skills, especially for key sectors including those identified in the new Industries. Moreover, HEI can find new area of research to attract funding bodies. In this way, they can contribute to the national economy.

### 2.2.19 Minimizing Cost of Higher Education Institutions

Higher education institutions' expenses have increased than before. On average, OECD countries spend USD 13,528 per student per year for all educational service, where UK spend about USD 16,000 and USA spend more than USD 25,000 (OECD, 2013). According to OECD (2013), expenditure per student by educational institutions is largely influenced by teachers' salaries, pension systems, instructional and teaching hours, the cost of teaching materials and facilities, the programme provided (e.g. general or vocational), and the number of students enrolled in the education system. Now a days student pay a lot, therefore their expectations become high and to meet their expectations HEI are spending a lot to increase support services to their students. It is observed that in the UK between 2004–05 and 2011–12 the number of full-time academic staff grew by 8220, and the number of part-time staff grew by 12510. Moreover, the statistics of 2011-12 on HEIs expenditure shows that HEIs spend about 56% which is a little over half of all expenditure in the HE sector for its academic and other staff costs and for other operating expenses HEIs spend about 37%, which is the second highest expenditure in the sector (Universities UK, 2013).

On the other hand, the funding bodies provide less than 40 per cent of revenue to most institutions (HEFCE, 2009b). At this financial restraint stage these institutions need to maintain their fiscal obligations with the limited budget, limited number of faculties and resources. Hence, minimizing cost of HEI becomes one of the major challenges in all HEIs (Beer, 2010).

HEIs are implementing a wide range of initiatives to enhance operational efficiency and reduce costs. In addition, JISC is working on how to minimize the cost of HEI with the help of information technology.

### 2.2.20 Increased Engagement with Industry, Business and Wider Community

Collaboration between higher education institutions and business is clearly linked to economic growth. The Higher Education Innovation Fund is distributing £150 million for 2010-11 for the development of a collaborative project with industry, business and wider community (HEFCE, 2010a; 2010b). HEFCE, QAA encourage HEI to increase relationship with industry to enhance the quality of learning opportunity engaging employers in designing courses and for student placement, which will further enhance students employability in this competitive era (QAA, 2008a; 2008b; HEFCE, 2010a; 2010b). According to Education and Skills Committee (2003) in the UK, HEI also need stronger links with business and economy for promoting economic development. JISC is also working on business and community engagement with HEI.

### 2.2.21 Higher Education Leadership and Management

HEI’s governing bodies are responsible for ensuring the effective management of the institution and for planning its future development (Prime Minister’s Strategy Unit, 2007; BIS, 2009). They are ultimately responsible for all the affairs of the higher education institutions. Generally, they are responsible for approving institutional mission and the strategic plan, financial solvency, resourcing policy, employment and Human Resource (HR) policy and strategy, estates policy, senior appointments and remuneration, audit, legal compliance, determining educational character and mission and so on. They are facing challenges to effectively manage the higher education institutions and hence, become one of the crucial challenges in HE (Hirsch and Weber, 1999; Education and Skills Committee, 2003; ACE *et al.*, 2006).

To cope with this challenge, institutions need better leadership who will be able to provide academic freedom and will be able to make collective decision with the new requirements that is the necessity to make and implement important and often unpopular decisions in a timely manner (Hirsch and Weber, 1999; Hanna, 2003). HEFCE aims to work in partnership with the HE sector to ensure that the HE system is run in the most effective and efficient way to secure its own long-term sustainability and to maintain its world class reputation for excellence. HEFCE have invested in improving leadership, governance and management in the higher education (HE) sector through the Leadership, Governance and Management (LGM) Fund.

### 2.2.22 Tenure

The rapidly changing world, the speed of knowledge creation, and economic pressures are causing higher education institutions to place greater emphasis on flexibility. Hence, tenure becomes another crucial and difficult issue in higher education institutions (West, 1999; Prime Minister's Strategy Unit, 2007). Higher education institutions must concentrate to effectively manage this issue for their greater interests. For example, they can replace resources at the expense of others while there is a need. Another example, senior faculty who are no longer productive can be replaced by hiring new faculty in an emerging discipline. However, at the same time, measures should be taken to offer alternative solutions for those losing tenure, like offering alternative occupation within or outside the institution or introducing a flexible age-of-retirement scheme (West, 1999).

## 2.3 Summary

Over the last decade, higher education sector has changed significantly around the world. In this chapter, we have presented twenty-two broad challenges facing higher education institutions namely: higher education leadership and management, higher education funding, widening participation, student retention, progression and completion, contribution to economy, assessment and feedback, plagiarism, group formation in learning and teaching, construction of personal and group knowledge, critical thinking and argumentation and so on.

A review of the literature suggests that student retention, progression and completion is one of the burning issues in the higher education sector around the world for many years. An analysis of student retention, progression and completion has a long tradition within the United States since late 1800's (Boston *et al.*, 2009). Moreover, Many theoretical models have developed to guide and assist in the understanding of the underlying reasons for early departure by students (Braxton, 2000). The UK higher education is increasingly focusing on student retention, progression and completion and it becomes one of the top agenda to be addressed in many UK higher education institutions. Moreover, data related to address student retention, progression and completion are widely collected and stored by higher education institutions.

This challenge and tensions provide the footing need to validate the need for this doctoral research in systematic analysis of student retention progression and completion. The next chapter explore in more detail the issue of student retention, progression and completion.

# Chapter 3: Student Retention, Progression and Completion

Student retention, progression and completion remains a central policy issue demanding active consideration by policy makers and those engaged in higher education across the globe. This chapter explores student retention, progression and completion on seven fronts. First, it discusses the definition of student retention, progression and completion. Second, it discusses how higher education institutions monitor student retention, progression and completion. Third, it discusses why it is important and what uses do the higher education institutions make of the data. Fourth, it presents retention, progression and completion related data and datasets. Fifth, it presents five well-known retention theories and discusses some retention studies based on the most accepted and popular Tinto's student integration model. Sixth, it discusses early prediction of students' performance/marks in monitoring poor performing students to retain and complete their study successfully. Finally, it discusses issues of traditional survey-based retention model.

## 3.1 Search Methods and Information Sources

To narrow down the online search from the voluminous amount of research relating to retention, the following keywords were used to search in the Google scholar and ISI Web of Knowledge: student retention theory, retention models, student progression, student completion, student attrition, student persistence, retention interventions, student retention risk factors, students' mark/performance prediction and identifying poor performing students. Furthermore, all articles' bibliographies were searched for additional resources. In addition, the following specific journals were utilized: Journal of College Student Retention: Research Theory and Practice, Journal of Higher Education and Colleges and Universities, Research in Higher Education. Also, University of Southampton's library resources were accessed. In addition, HEA, HEFCE, NAO, QAA, HESA's reports and publications were used as an information source because they work actively for UK higher education and provide reliable information and statistics regarding higher education in the UK.

## 3.2 What is meant by Student Retention, Progression and Completion?

Student retention is a complex issue and there is no single definition for student retention. It is defined in various ways in different literature. Moxley *et al*. (2001, p. 37) define student ***retention*** as

> *"the process of helping students to meet their needs so they will persist in their education toward the achievement of educational aims they value. Retention can achieve this through the assembling of supports that enable students to be successful and the lowering or elimination of those factors that can disrupt the students' education and that can ultimately result in their failure to achieve these educational aims they want"*

which is a widely used definition in many studies. In his review of the research literature on student retention, Jones (2008) defines student ***retention*** as

> *"the scope to which learners persist within a higher education institution, and complete a programme of study in a pre-determined time-period".*

In the UK, the Higher Education Funding Council for England (HEFCE[21]) uses two main measures for retention: completion rate and continuation rate. The ***completion*** rate is the proportion of students who start in a year and continue their studies until they obtain their qualification, with no more than one consecutive year out of higher education and the ***continuation*** rate is the proportion of the annual intake of new students who return to higher education in the subsequent year. Whereas, ***progression*** rate is the proportion of students who move to the next level of a programme of study at the end of

[21] http://www.hefce.ac.uk/

an academic session (NAO, 2007). For example, progression is the number of students who completed their first year and go to the second year.

Therefore, in this study of student retention, progression and completion, the definitions of the HEFCE will be used. Hence, completion is the number of students who complete their programme within a pre-determined time period, continuation is the number of students who return to higher education in the following year of their study and progression is the number of students from one academic year to the next; for example, moving from year 1 to year 2.

## 3.3 How do Universities Monitor Student Retention, Progression and Completion?

Each year institutions have to return data, which relate to the number of students that they have enrolled, and data on the progress of students who are already enrolled to the Higher Education Statistics Agency (HESA[22]). Based on those data, since 2004, HESA have been publishing annual performance indicators (PI). Before 2004, the English Funding Council HEFCE published Performance Indicators from 1999 to 2003 for individual universities. The indicators are intended to provide reliable and comparable performance information of institutions about widening participation, student retention, learning and teaching outcomes, research output and employment of graduates, which is helpful for a range of users, including prospective students, universities and the funding council. The publication of performance indicators provides targets for universities to improve student retention (The House of Commons, 2008). In addition, since 2005, a national survey of university students (NSS[23]) has been conducted which measures students' satisfaction in teaching and learning, assessment and feedback, academic support, organization and management, learning resources and personal development. This survey's results are published on Unistats[24] web site, so that potential students can make informed judgements of their potential success based on the track record of the institutions to which they are applying. The results are also worthwhile to universities to facilitate best practices and to

[22] http://www.hesa.ac.uk/
[23] http://www.thestudentsurvey.com/
[24] http://unistats.direct.gov.uk/

enhance the student learning experience in areas where they found their students are dissatisfied.

Higher education institutions (HEI) can use performance indicators together with the student satisfaction information to improve student retention, progression and completion, as this information affect universities' reputations and numbers of student applications. (NAO, 2007; The House of Commons, 2008).

HEI have introduced some different ways to improve student retention, progression and completion. They are recognising that understanding the reasons of student non-completion is vital for an institution to increase student success. A number of HEI have taken steps to research the reasons for non-completion and then develop ways to increase retention (HEFCE, 2001). A number of theories have been developed by researchers on student retention over many years. Theory and research help institutions to encourage student success.

The first and most commonly used model in the student retention literature is Tinto's model (1975; 1987; 1993), proposing a multivariate model of student retention in universities and colleges to explain early student departure; where the likelihood of a student withdrawn is seen as being determined by individual attributes, familial attributes, prior qualifications, social integration, academic integration, individual commitment, institutional commitment and external family and societal factors taking place during the course of study. Tinto claims that students who are highly integrated academically are more likely to continue and complete their degrees and students who have more friends at their institutions, have more personal contact with academics, enjoy being at the institution, are likely to make the decision to remain in their institutions. Since Tinto's models gained prominence, there have been various studies of the social and academic integration approach (Napoli and Wortman, 1996; Sullivan, 1997; Thomas, 2002; Pascarella and Terenzini, 2005).

Yorke (1999) and Davies and Elias (2003) conducted two large quantitative higher education studies in the UK to examine the reasons students withdraw early from their courses. Many retention studies recognise a variety of personal reasons and institutional factors affecting withdrawal from courses. For example, Yorke (1999) reported wrong choice of field of study, financial problems, academic difficulties, poor quality of the student experience, unhappiness with the social environment as the most important factors to withdraw from courses. Davies and Elias (Davies and Elias, 2003) also found that wrong choice of course, financial problems, personal problems, academic difficulties and wrong choice of institutions are the most common reasons to withdraw from the courses. Recently, the NAO (2007) also stated that students withdraw from courses for a variety of interrelated reasons and mentioned some common reasons for students' withdrawal. For example, NAO also mentioned wrong choice of course, financial reasons, lack of integration, lack of preparedness, dissatisfaction with course/institutions, personal reasons and to take up a more attractive opportunity.

The first year of study is recognized as a key stage, as during this period a new student is most likely to dropout from HEI (Thomas *et al.*, 1996; Tinto, 1998; Yorke, 1999; HEFCE, 2001; Harvey *et al.*, 2006). Yorke (1999) noted about one third of students and Thomas *et. al*, (1996) noticed about 77% of students withdraw from their courses during their first year. Beyond this stage, it is important that students are provided with and have access to a high level of continuous support (Harvey *et al.*, 2006). Some non-completion of courses is unavoidable and should not be viewed as failure of the student, tutor or college. However, a lot of non-completion is preventable and it is the responsibility of the universities/colleges to help to retain their students (Yorke, 1999).

There is sizeable literature on support services to improve student retention, much of which outlines good practice and the need for appropriate and integrated interventions. HEI follow these and set their own retention strategies to achieve high retention rates. Some have developed ways based on the 'student life-cycle' introduced by HEFCE (2001). The NAO (2007) reports some of the important activities higher education institutions provide to improve student retention, such as specialist support services to students, financial support through bursaries and hardship funds, flexible learning options to fit personal circumstances and personal tutoring systems (Sullivan, 1997) to individual students for extra support and facilities to improve their chances of success. Both academic and pastoral support is important to enhance the student experience. Universities provide pastoral support and counselling services for students in different ways by providing personal tutor, and personal mentor.

Apart from these, Dodgson and Bolam (2002) report that HEI also arrange pre-entry activities including open days, pre-entry information and advice, and guidance to all prospective students. Some HEI provide prospective students with the opportunity to attend an access course or attend a summer school. HEI organize induction or orientation days to ensure that students are familiar with university procedures, practices, customs and expectations (Edward, 2003; Yorke and Thomas, 2003). Most HEI have some form of attendance policy, including arrangements for contacting students who are absent from a certain number of lectures or tutorials (NAO, 2002). Universities have started to offer career advice and guidance to their students and also include employability skills and engage employers in designing the curriculum to support student employability (NAO, 2002), as many students enter university because they have specific career paths in mind. For example, University of Southampton's Career Destination actively works on enhancing student employability by providing advice and information to students, organizing career fairs, seminars and workshops. Moreover, HEI use statistical data to improve student retention, progression and completion (NAO, 2007; The House of Commons, 2008).

Universities collect and use management information on withdrawal rates and reasons for leaving to produce regular reports for planning and decision making and tracking the performance of students to highlight those that may need more support (NAO, 2007; QAA, 2008c; The House of Commons, 2008). However, there is no reliable national data on reasons for leaving because universities do not always collect the information when students leave courses. The Funding Council advise that all universities should establish reasons for leaving; for example, through exit interviews, and should have systems to identify and investigate trends in withdrawal and take necessary steps as required (The House of Commons, 2008) .

Currently, HEI are taking a number of actions to improve student retention, progression and completion as we discussed earlier. Figure 3.1 shows these activities in three key stages i.e., Pre-entry activities, Fair admission process and Supportive student life. Therefore, it is easy to recognise in which stage students need what type of support.

## 3.4 Why do HEI Monitor Student Retention, Progression and Completion and What Use do HEI Make of the Data?

Student retention, progression and completion are very significant to HEI. Student retention has been the focus of research on higher education, not least due to efforts to establish a benchmark indicator of institutional performance and to gain a better understanding of enrollment-driven revenue streams (Herzog, 2005).

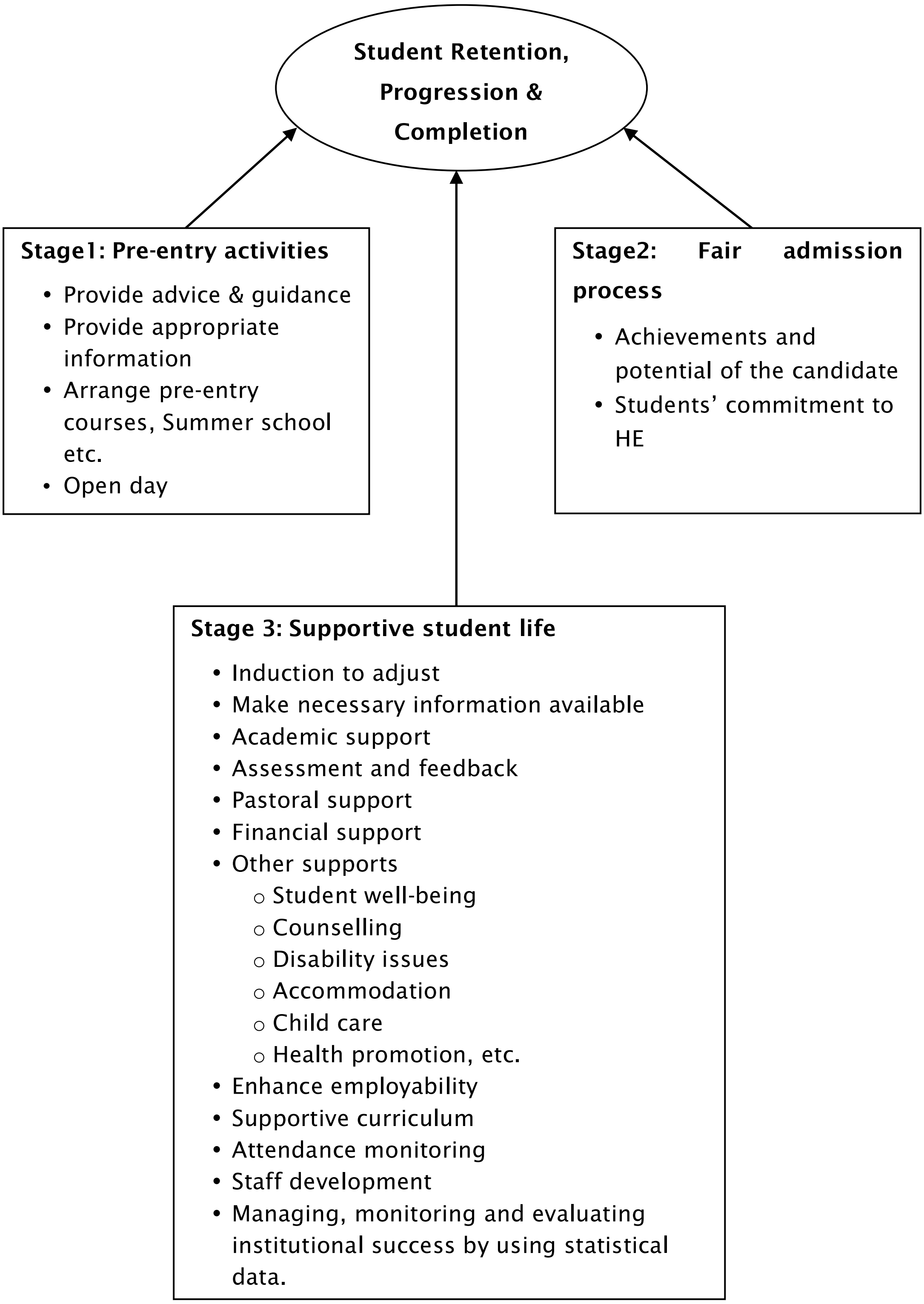
The National Audit Office (NAO) in (2002) stated that there is currently much interest in not just access to higher education, but student success too. In the UK, the Higher Education Academy (HEA[25]) is an organisation established to enhance the student experience at universities in the UK. According to its former chief executive, Paul Ramsden,

> *"Student satisfaction and retention are at the heart of an educational institution. A high quality student experience is the hallmark of excellent higher education".*

Also, Bailey and Borooah (2007) noted two main reasons for being concerned about the low rates of student retention in higher education. One of which is the associated wastage of resources and the other one is attracting prospective students. Furthermore, the Parliamentary Select Committee on Education and Employment also stated the same reason to concern, which is clear from the statement,

> *"Increasing non-completion rates could undermine success in opening higher education to a broader spectrum of the population, put off potential students, and cause institutional instability." (Parliamentary Select Committee on Education & Employment, 2001).*

[25] http://www.hefce.ac.uk/news/hefce/2009/nss.htm



**Student Retention, Progression & Completion**

**Stage1: Pre-entry activities**

- Provide advice & guidance
- Provide appropriate information
- Arrange pre-entry courses, Summer school etc.
- Open day

**Stage2: Fair admission process**

- Achievements and potential of the candidate
- Students' commitment to HE

**Stage 3: Supportive student life**

- Induction to adjust
- Make necessary information available
- Academic support
- Assessment and feedback
- Pastoral support
- Financial support
- Other supports
  - Student well-being
  - Counselling
  - Disability issues
  - Accommodation
  - Child care
  - Health promotion, etc.
- Enhance employability
- Supportive curriculum
- Attendance monitoring
- Staff development
- Managing, monitoring and evaluating institutional success by using statistical data.

Figure 3.1 The process of improving student retention, progression and completion as suggested by the literature review.

The UK Quality Assurance Agency (QAA) advises HEI to take student retention seriously to improve as it is a key performance indicator in educational quality and HEFCE's provision for funding institutions is based on the numbers of students completing their year (QAA, 2008c; HEFCE, 2010a). Universities must focus on enhancing student experience to improve student retention. Otherwise, universities can lose funding if they retain fewer students than expected (The House of Commons, 2008). On the other hand, high student retention rate has a positive effect on student employability, as employers think that graduates are well prepared for work, and widening participation, through attracting international students (Prime Minister's Strategy Unit, 2007). Hence, HEI are more conscious about student retention, progression and completion.

Universities collect and use data for planning and decision making and tracking the performance of students to highlight those that may need more support (NAO, 2007; QAA, 2008c; The House of Commons, 2008). For example, from the institutional data, HEI can know in which area their students are dissatisfied and they can focus on that area to improve and plan for budget accordingly. HEI can also monitor student achievements and can identify those students who are at-risk and can provide appropriate supports to help them succeed. Moreover, HEI use these data to monitor their achievements internally, to increase their performance, and to help them decide on the most effective way of spending their money in order to achieve the best for their student experience. From the literature, it is suggested that institutions should do more with the data they collect that relates to the first year of study (Johnston, 2001; Harvey *et al.*, 2006).

## 3.5 Student Retention, Progression and Completion Related Datasets

Based on the literature, a list of variables related to student retention, progression and completion are collected and categorised into five terminologies: students' individual attribute, academic preparedness, academic variable, support and institutional variable. These are presented in *Appendix* C. In addition, a summary of the possible institutional datasets related to student retention, progression and completion are provided:

- Dataset that holds information about student records, for example, student background, examination results and course enrollment. Student records system is the most important data source.
- Dataset that holds logs of students' access to online resources such as e-book and online journal or etc.
- Dataset that records student attendance.
- Dataset that holds logs of students' submission of their assignment/homework.
- Dataset which provides information about student library loan history with details about books borrowed and how often the student borrows books.
- Payroll system of the university that hols information about student's working information in the university, for example, the student works or not and if works, the number of hours the student works.

Institutions can integrate these datasets for analysing data to improve student retention, progression and completion.

## 3.6 Student Retention Models

Student retention in higher education has been the subject of an enormous amount of research over seven decades (Braxton, 2002). Several student retention models have been developed by researchers to identify and analyze the factors affecting student retention. Researchers have studied student retention in higher education from five theoretical perspectives: psychological, social, economic, organizational, and interactional.

The psychological perspective focuses on individual personality attributes and views in retaining students. The key theories in this category are Astin's (1984) Student Involvement theory and Bean and Eaton's (2000) Psychological theory. In contrast, the social perspective focuses not on the individual, but rather on social forces that are external to the higher education institution such as social status, race, prestige and opportunity (Tinto, 1993). The economic perspective focuses on the individual finance and financial aid that affects student retention (Tinto, 1993). The organizational perspective is concerned with the impact of organizational factors such as bureaucratic structure, size, faculty-student ratios, and institutional resources and goals on student retention. Organizational theories are useful in explaining student retention between higher education institutions. However, they are less useful in explaining student retention within institutions (Tinto, 1993). The key theory in this category is Bean and Metzner's (1985) Student Attrition Theory. The interactional perspective focuses on the influence of the interaction of individual and environmental factors on student retention. Tinto's (1975; 1993) Student Integration Theory is the key theory in this category.

This review of the literature examines five of the most widely tested theories of student retention. These are Tinto's (1975, 1993) Student Integration Theory, Pascarella's (1980) Attrition Theory, Astin's (1984) Student Involvement Theory, Bean and Metzner's (1985) Student Attrition Theory and Bean and Eaton's (2000) Psychological Attrition Model.

### 3.6.1 Tinto's (1975, 1993) Student Integration Model

Tinto's (1975) Student Integration model is the most widely discussed and most researched model of student retention. Berger and Braxton (1998) have stated that Tinto's integration model has been the focus of much empirical research and has near-paradigmatic status in the study of the college student departure. Tinto's model is based upon Durkheim's theory of suicide and Spady's theory of departure. Tinto's model is a longitudinal process and regards student retention as the degree to which a student becomes integrated into the social and academic life of the university (Tinto, 1993; Rendon *et al.*, 2000). Academic integration is defined as student's perceived academic performance and intellectual development while social integration is defined as the quality of a student's relationships with both the peer group and the faculty (Pascarella and Terenzini, 1980). Tinto (1993) points out that both types of integration do not need to be equal, but some level of academic and social integration must occur in order for students to persist in the university. In addition, Tinto also points out that both types of integration may have a reciprocal relationship. For example, if a student is very connected in the academic life by spending too much of his time in study then the student may have a lack of social integration in the university. As a result, this may have a negative consequence with regard to student retention.

According to Tinto's (1975) theory, students enter the university with a set of background characteristics including family backgrounds (e.g., family social status, parental formal education, and parental expectations), individual attributes (e.g., gender, race, age, and academic aptitude), and pre-college schooling (e.g., high school achievement, academic course work). These background characteristics combine to influence the initial goal and institutional commitments that the student brings to the university environment. Goal commitments represent the degree to which the student is committed, or motivated, to get a university degree. Institutional commitments represent the degree to which the student is motivated to graduate from a specific university. These commitments change during the student's time at the university as a result of the degree of integration into the academic and social systems of the university. In turn, these two types of integration lead to new levels of goal and institutional commitments. In addition, the student's initial goal and institutional commitments influence their later goal and institutional commitments. Finally, the later goal and institutional commitment determines whether or not the individual decides to drop out from college (Tinto, 1975).

In addition to Durkhiem's theory, Tinto also incorporated Van Gennep's theory about rites of passage to explain his model. From Van Gennep, Tinto included the concepts of separation, transition, and incorporation.

The first stage of the college student experience is separation. It requires students to disassociate themselves physically and socially from their previous communities such as high school friends, family, and place of residence. These previous communities often have different values, norms, and behavioral and intellectual styles than university. As a result, there must be some degree of transformation and possibly rejection of the norms of previous communities in order for the students to successfully integrate into the norms of the university community. Students who attend a local, non-residential university may not have to disassociate themselves completely from previous communities but they may not be able to fully integrate academically and socially into the new university community (Tinto, 1988; 1993).

The second stage of the student experience is transition. It comes either during or after the separation stage. It is the stage where students find themselves separated from their previous communities but have not yet fully adapted to the university community. Many students voluntarily withdraw from university during this stage because they cannot cope with the many stresses of transition. However, a student's goals and institutional commitment play an important role in this stage. If the student is committed to the goal of education and to the university, then he can overcome the stresses of transition (Tinto, 1988).

The last stage is incorporation. It can only happen when students have passed through the stages of separation and transition, which tend to occur early in the student's experience. In this stage, the students are expected to become integrated or engaged into the university community. However, unlike incorporation into traditional societies, students are often not provided with formal rituals and ceremonies to engage them to the university community. It is important for the university to provide a variety of formal and informal mechanisms to engage students to the university community, including residence hall associations, student organizations, extracurricular programs, and faculty lectures (Tinto, 1993).

Tinto modified his original model in 1993 with the addition of two constructs or factors: External Commitments and Intentions. According to Tinto (1993), a student's intentions have a direct influence on their goal and institutional commitment, which both directly influence student retention. External commitments such as families, neighborhoods, peer groups and work environments can also have a direct influence on student's initial goal and institutional commitments. Figure 3.2 presents Tinto's modified model.

### 3.6.2 Pascarella's (1980) Student Attrition Model

Pascarella's (1980) Attrition model is based on both Spady's (1970) and Tinto's (1975) model. His model emphasises the informal interactions between student and faculty as being important in students' educational outcomes and retention. Pascarella's model is longitudinal. According to Pascarella (1980):

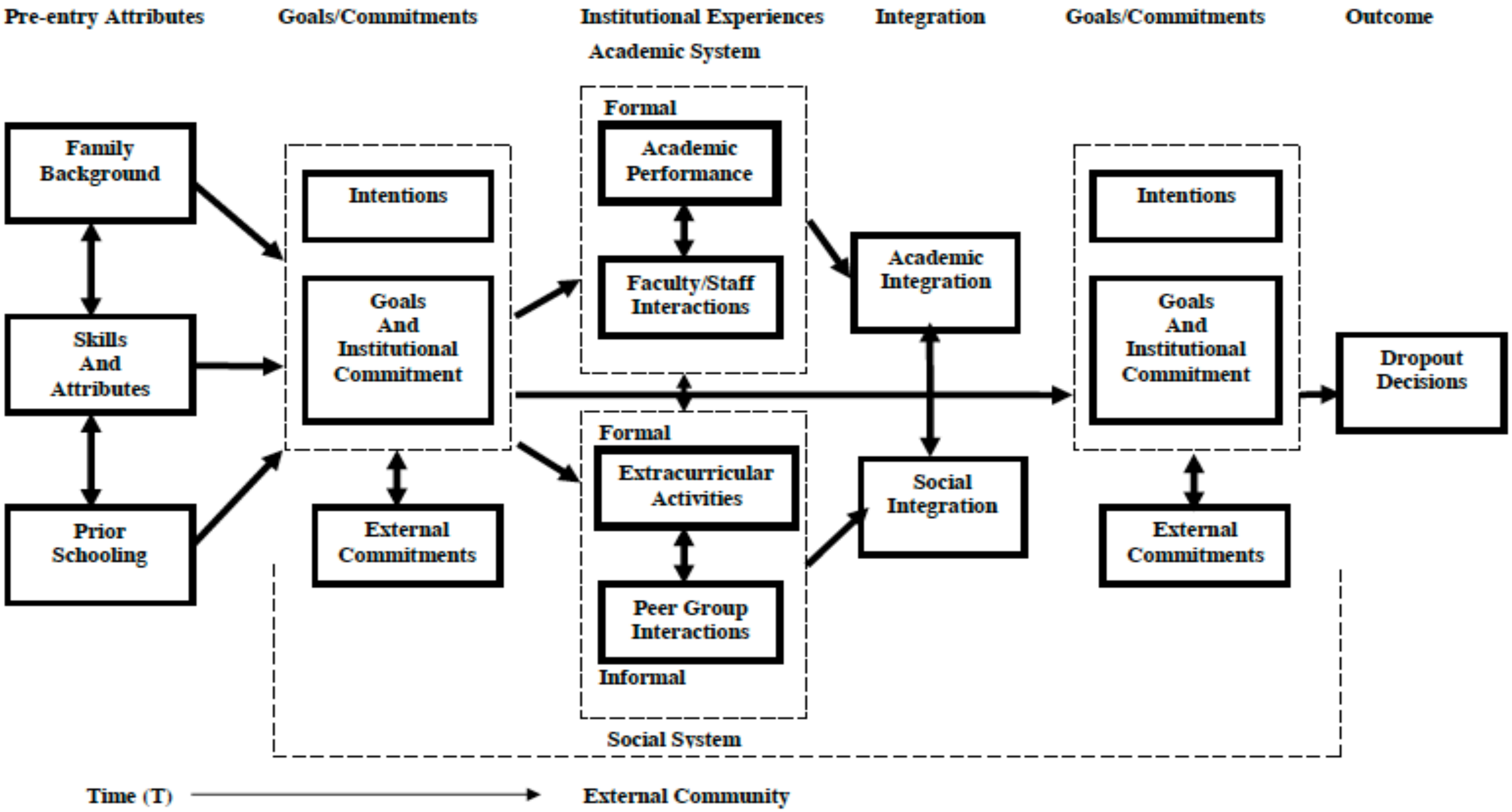


Figure 3.2 Tinto's (1993) Student Integration Model (Tinto, 1993, pp. 114. Reproduced with permission).

> "*In order to understand the unique influence of student faculty non-classroom contact on educational outcomes and institutional persistence, it is necessary to take into account, not only background characteristics which students bring to college, but also actual experiences of college in other areas, as well as salient institutional factors.*" (pp.568)

According to Pascarella's theory (1980), presented in Figure 3.3, student characteristics and institutional characteristics influence each other and the three independent variables. The three independent variables include informal contact with faculty, other college experiences, and educational outcomes. The three independent variables reciprocally affect each other so that a problem in one area may affect another area. Only educational outcomes have a direct influence on student retention decisions. All other variables affect the persistence/withdrawal decision indirectly through their affect on educational outcomes.

### 3.6.3 Astin's (1984) Student Involvement Model

Astin's (1984) Student Involvement model simply states that students learn by becoming involved/engaged. It emphasizes that the factors important to student involvement/engagement are important to stay enrolled in the university. Astin (1984) defined student involvement as:

> "*The amount of physical and psychological energy that the student devotes to the academic experience. Thus a highly involved student is one who, for example, devotes considerable energy to studying, spends much time on campus, participates actively in student organizations, and interacts frequently with faculty members and other students."* (Astin, 1984)

This means, the students who are involved/engaged give significant energy to academics, spend time on campus, participate actively in student organizations and activities, and interact with faculty. On the other hand, uninvolved students neglect their studies, spend little time on campus, abstain from extracurricular activities, and rarely initiate contact with faculty or other students.

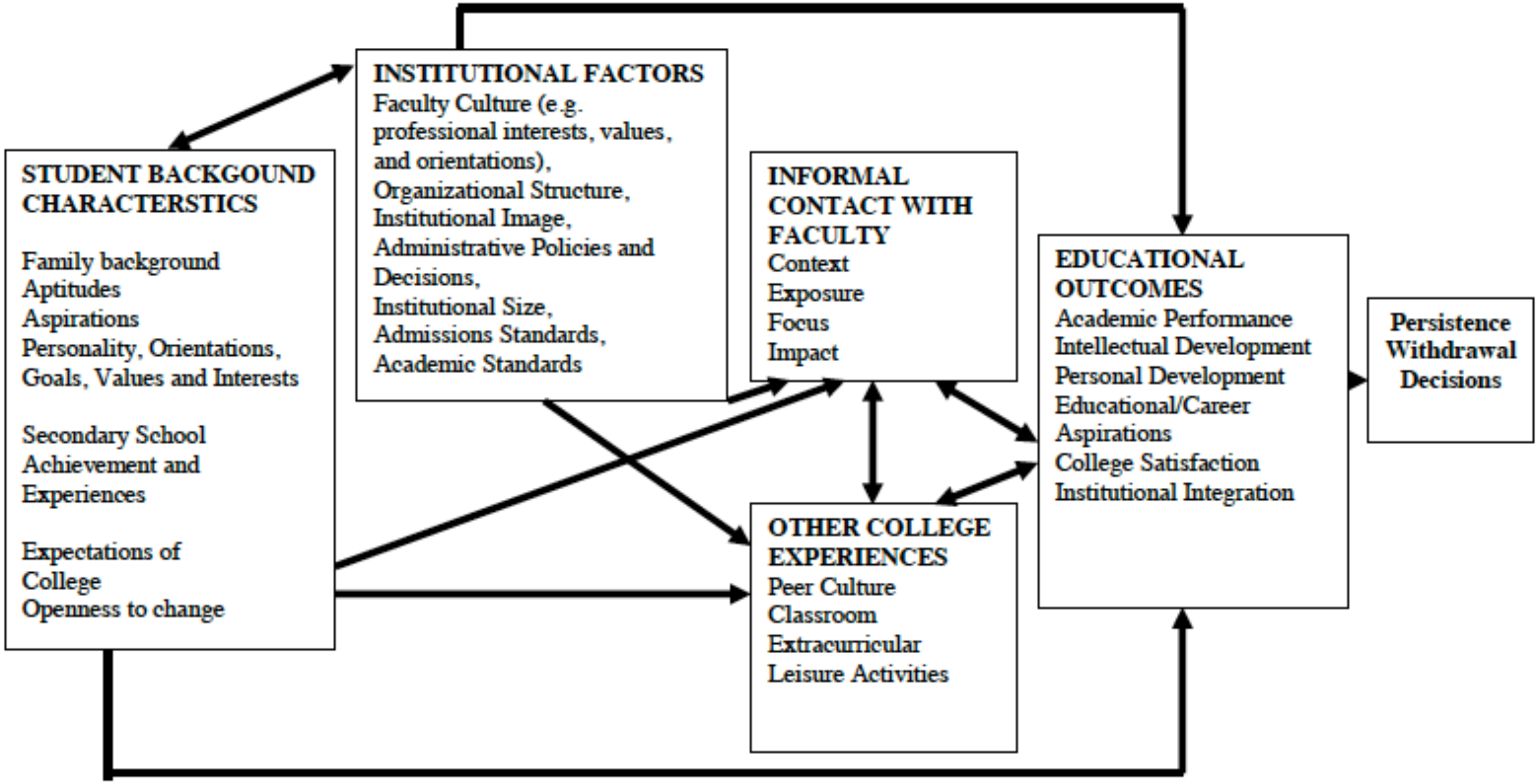


Figure 3.3 Pascarella's (1980) Student Attrition Model (Pascarella, 1980, pp. 569. Reproduced with permission).

Astin's (1984) student involvement theory contains five basic postulates:

- Involvement/engagement requires the investment of energy (physical and psychological).
- Students invest varying amounts of energy in the tasks facing them.
- Involvement/engagement has both quantitative (e.g., the numbers of hours a student spends studying) and qualitative (e.g., the amount of learning that takes place during study time) features.
- The amount of student learning and development is directly proportional to the quality and quantity of involvement/engagement.
- The education effectiveness of a policy or practice depends on its ability to stimulate students' involvement/engagement.

### 3.6.4 Bean and Metzner's (1985) Student Attrition Model

Bean and Metzner's (1985) Student Attrition model is based on organizational turnover theory and attitude-behaviour interactions theory. It emphasizes that student decisions to leave university are synonymous with adult decisions to leave the workplace. Bean and Metzner developed this model for non-traditional students. They contend that the student retention models developed by Astin and Tinto relied too heavily on socialization to explain retention and did not take into account the external factors affecting non-traditional students who have fewer opportunities for social integration. They define non-traditional student by age, residence, and attendance. According to Bean and Metzner (1985):

> "*A nontraditional student is older than 24, or does not live in a campus residence (e.g., is a commuter), or is a part-time student, or some combination of these factors; is not greatly influenced by the social environment of the institution; and is chiefly concerned with the institution's academic offerings (especially courses, certification, and degrees).*" (pp. 489)

Bean and Metzner's (1985) Student Attrition Theory, presented in Figure 3.4, posits that four sets of variables influence student retention:

- Academic variables, measured by grade point average.
- Student's intention to leave, which is expected to be influenced primarily by psychological outcomes (institutional quality, satisfaction, goal commitment and stress) and academic variables.
- Background and defining variables (primarily high school performance and educational goals).
- Environmental variables such as finances, hours of employment, family responsibilities and opportunity to transfer.

Bean and Metzner find that environmental variables are more important than academic variables for non-traditional students:

> "*When academic variables are good but environmental variables are poor, students should leave school, and the positive effects of the academic variables on retention will not be seen. When environmental support is good and academic support is poor, students would be expected to remain enrolled, the environmental support compensates for the low scores on the academic variables." (Bean and Metzner, 1985).*

Similarly, they find that psychological variables are more important for nontraditional students than academic variables. In other words, if scores on both variables are high, students are more likely to persist and if both are low, the students are more likely to drop out. If the psychological variables are low and the academic variables are high, the students are more likely to drop out. Conversely, if the psychological variables are high and the academic variables are low, the students are more likely to persist.

### 3.6.5 Bean and Eaton's (2000) Psychological Attrition Model

The primary theme of their model is that student departure is the result of the premeditated intention to leave. As described by Bean in (Seidman, 2005), "Intention is based on prematriculation attitudes and behaviors that

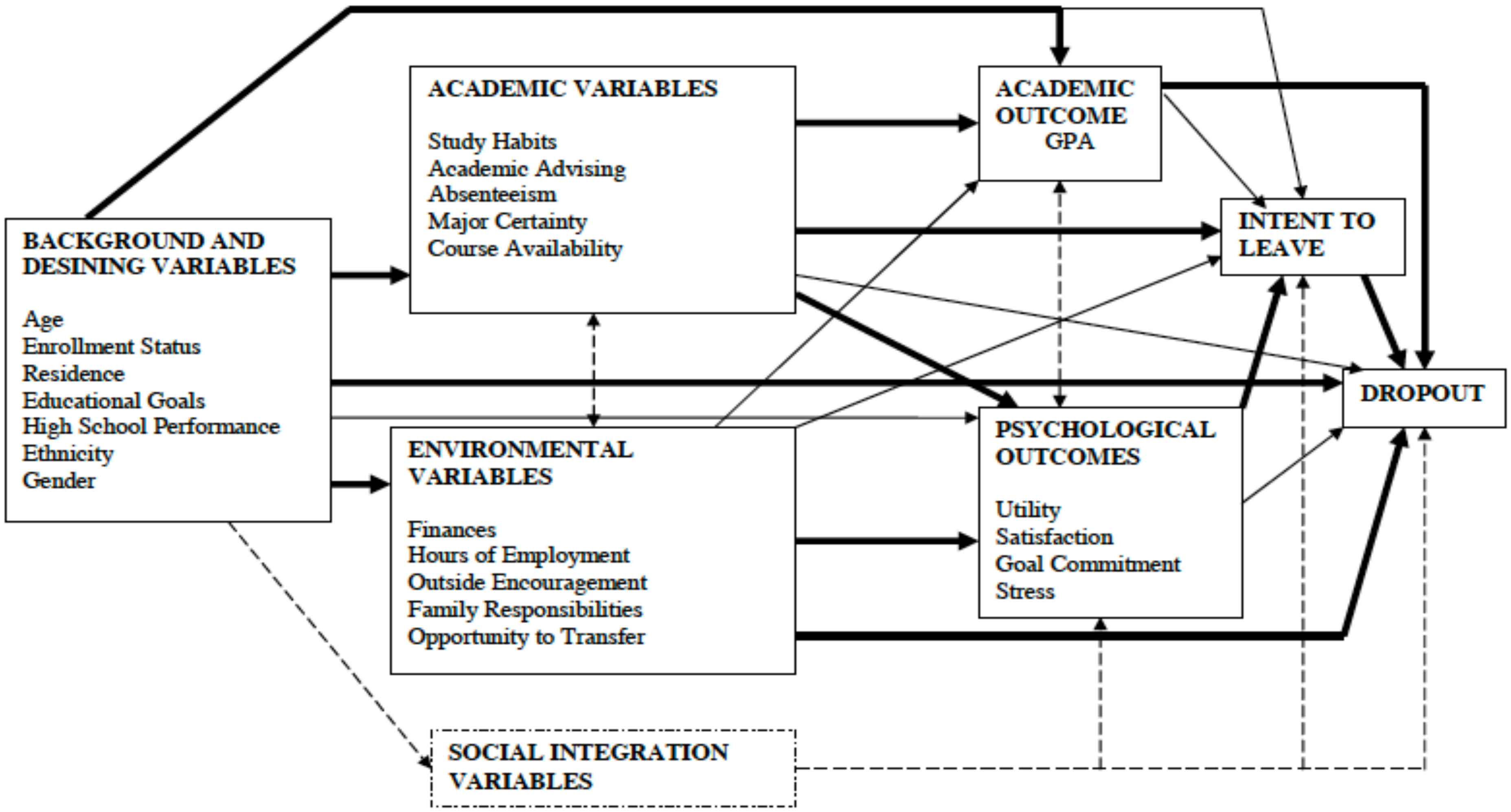


Figure 3.4 Bean and Metzner's (1985) Student Attrition Model (Bean *et al.*, 1985, pp. 491. Reproduced with permission).

affect the way a student interacts with the institution. On the basis of this interaction, the student develops attitudes towards their experiences and norms related to student behavior." As with Tinto's (1993) model, Bean's model is longitudinal in nature and reflects the student's attitudes and behaviors as they navigate the educational experience. The model is also summarized by Bean and Eaton (2002) as follows:

> *"An individual enters an institution with psychological attributes shaped by particular experiences, abilities, and self-assessments. Among the most important of these psychological factors are self-efficacy assessments ("Do I have confidence that I can perform well academically here?"); normative beliefs ("Do the important people in my life think attending this college is a good idea"); and past behavior ("Do I have the academic and social experiences that have prepared me to succeed in college?")." (p. 75)*

The student then interacts with the institution (its bureaucratic, academic, and social realms) while continuing to interact with people (parents, spouses, employers, and old friends) who are outside of the institution. These interactions include staff from various departments, their faculty, both inside and outside the classroom, and also with other students.

Figure 3.5 presents Bean and Eaton's (2000) model. The model depicts the student's psychological processes as they interact with and respond to their environment. Similarities can be seen with Tinto's (1993) model, such as the precollege attributes, which the student brings with them to college and which informs their attitudes and predisposition to stay enrolled or drop out.

Five of the most widely tested student retention models were reviewed. These were Tinto's (1975; 1993) Student Integration Model, Pascarella's (1980) Attrition Model, Astin's (1984) Student Involvement Model, Bean and Metzner's (1985) Student Attrition Model and Bean and Eaton's (2000) Psychological Attrition Model. In reviewing the foundational student retention literature, it is found that Tinto's (1975; 1993) student retention model is one of the most studied and dominant in the field of higher education. Berger and Braxton (1998) have stated that Tinto's integration model has been the focus of much empirical research and has near-paradigmatic status in the study of student

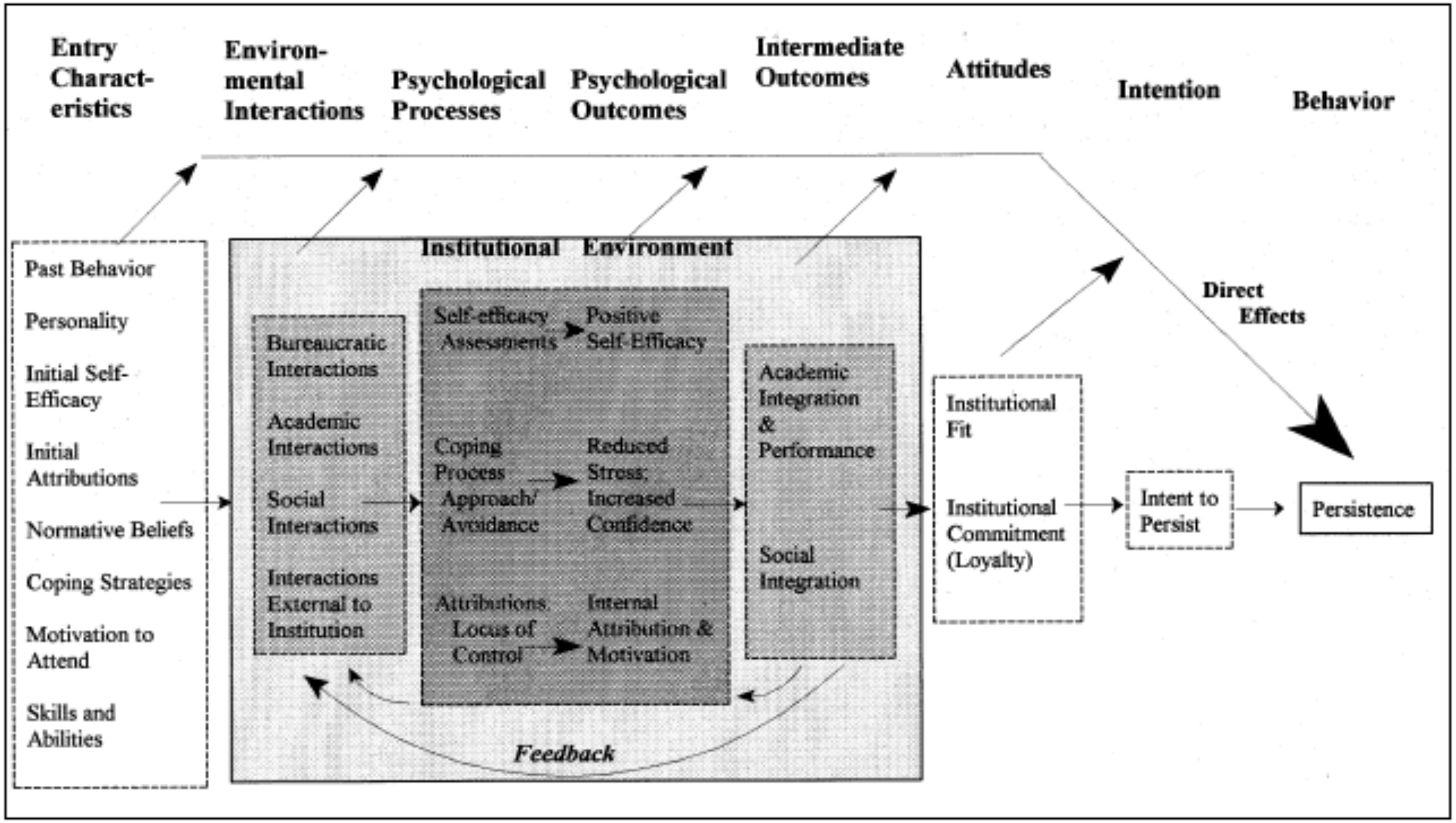


Figure 3.5 Bean and Eaton's (2000) Psychological Model of Student Retention (Bean and Eaton, 2002, pp. 76. Reproduced with permission).

retention. Many studies by other researchers have employed Tinto's (1975) model as a starting point in their investigations of student retention. The following section will review studies based on Tinto's model.

## 3.7 Student Retention Studies based on Tinto's Integration Model

In this section, studies testing Tinto's model is reviewed. There are many variations in how the researchers have tested/studied Tinto's model: tested the whole model, tested the whole model with the addition of other constructs, tested parts of the model and tested parts of the model with the addition of other constructs. According to Tinto's model, student retention is measured by four groups of variables: students' background characteristics/pre-entry attributes, current academic experience, social and academic integration, and goal and institutional commitment. Researchers used various scales to measure students' background, social and academic integration, and goal and institutional commitment. A most popular scale to measure Tinto's all construct is Pascarella and Terenzini's (1980) Institutional Integration scale (IIS). The IIS has been used in various forms in the research (Nora *et al.*, 1990; Berger and Milem, 1999b; Berger and Braxton, 2000). Modifications to the scale have been made in an attempt to adapt the scale to match particular settings or populations. Table 3.2 shows IIS to measure all the constructs of Tinto's model.

In the following, we will discuss many of the studies that support the variables that make up Tinto's model as well as the studies that directly tested Tinto's model.

### 3.7.1 Background characteristics/Pre-entry attribute

According to Tinto's model, student's background variable or pre-entry attribute is broken down into the following three sub-areas: individual attribute, family background and prior schooling. Individual attribute is defined by age, gender, race/ethnicity, accommodation, residence, study field/major, whether the student is the first member of his/her family to attend university, standardized test scores and source of tuition fee. Family background is included parents' annual income and parent's education level. Prior schooling is defined by a student's high school scores and high school class rank.

Table 3.2: IIS to measure all the constructs of Tinto’s model

| |
|---|
| **Academic and Intellectual Development**<br>I am satisfied with the extent of my intellectual development this year.<br>My academic experience this year has had a positive influence on my intellectual growth and interest in ideas.<br>I am satisfied with my academic experience at Cenfral Christian University this past year.<br>My interest in intellectual ideas and intellectual matters has increased this year.<br>I am more likely to attend a cultural event (for example a concert, lectiire or art show) now than I was a year ago.<br>I have performed academically as well as 1 anticipated I would. |
| **Faculty Concern for Student Development and Teaching**<br>Few of the faculty members that I have had contact witii this year are genuinely interested in students.<br>Few of the Faculty members I had contact with this year are genuinely outstanding or superior teachers.<br>Few of the CCU faculty members I have had contact with this year are willing to spend time outside of class to discuss issues of interest and importance to students.<br>Most of tiie CCU faculty members I have had contact with are interested in helping students grow in more than just academic areas.<br>Most faculty members I have had contact with this year are genuinely interested in teaching. |
| **Peer Group Interaction**<br>The student friendships I have developed this past year have been personally satisfying.<br>I have developed close personal relationships with other students.<br>My interpersonal relationships with other students have had a positive influence on my personal growth, values and attitudes.<br>My interpersonal relationships with other students have had a positive influence on my intellectual growth and interest in ideas.<br>It has been difficult for me to meet and make friends with other students.<br>Few of the CCU students I know would be willing to listen to me and help me if I had a personal problem.<br>Most students at CCU have values and attitudes, which are different from my own. |
| **Interactions with Faculty**<br>My non-classroom interactions with faculty this year have had a positive influence on my personal growth, values and attitudes.<br>My non-classroom interactions with CCU faculty members have had a positive influence on my intellectual growth and interest in ideas.<br>My non-classroom interactions with faculty this year have had a positive influence on my career goals and aspirations.<br>This past year, I have developed a close personal relationship with at least one faculty member. |
| **Institutional and Goal Commitments**<br>It is important for me to graduate from college.<br>I made the right decision in choosing to attend this institution.<br>It is not important for me to graduate from this institution.<br>I have no idea at all what I want to major in.<br>Getting good grades is not important to me.<br>I am confident that I made the right decision in choosing to attend this institution. |

Past research on retention have been shown conflicting result considering background characteristics in predicting student retention. Pascarella *et al.* (1983) stated that "*students' background 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 their study, it was found that background characteristics made the largest significant contribution in predicting student retention. However, background characteristics were found statistically non-significant in many studies (Terenzini and Pascarella, 1978; Fox, 1986). Some of the variables in this category are discussed below have found a direct relationship between certain pre-entry attributes and persistence/withdrawal.

## Gender

Past research on retention differences between men and women have yielded conflicting results. Ramist (1981) suggests that there are no drop-out differences between men and women. Hilton (1982) also reported little if any difference between male and female drop out rates using data collected from a national longitudinal study on retention. Murtaugh *et al.* (1999) and Stage (1988) also agreed with Ramist's (1981) findings and found that gender had no effect on retention. Feldman (1993) discovered that gender, when tested by itself, is associated with persistence. However, when other variables were included, the effects of gender on persistence were non-existent. On the other hand, some researchers found gender differences on student retention. In his study of retention, Berger (1997) reported that gender differences on persistence. Randall (1999) study also supported Berger's finding as he found that female students were more likely than male students to re-enrol in their study. In addition, after examining cohort groups, Randall (1999) discovered that female students had higher graduation rates than male students. Elkins *et al.* (2000) also found gender effect on persistence. However, he reported that being a female student increases the likelihood of early departure from the institution. The conflicting results of the previously mentioned studies make it difficult to ascertain whether or not gender is directly related to student persistence.

## Ethnicity/Race

Significant attention has been given to the retention rates of minority students throughout the past decades. The results of many of these studies showed that minority students drop out at higher rates than majority students, For example, in USA minority students especially African Americans and Hispanics drop out at higher rates than white students. In a study of Maryland four-year public institutions, Randall (1999) found that the six-year graduation rate for the entire 1992 cohort was 56 percent, while the graduation rate of the 1992 African American cohort was 40 percent. Porter (1990) found that over a six-year period, African Americans and Hispanics were more likely to dropout than both Asian American students and white students. This evidence was confirmed in Carroll's (1989) study. Carroll (1989) reported that those students planning to go to college from the Class of 1980, by 1983, 56 percent of white students, 44 percent of Afiican American students, and 42 percent of Hispanic students had persisted. Feldman (1993) indicated that African American students were 1.75 times more likely to drop out than their white counterparts. Astin (1975) found that African American students are more likely to drop out predominantly from white colleges (where majority students are white) than African American colleges (where majority of the students are African American). Similar observation also found by Lenning *et al.* (1980) and reported that Spanish-speaking students drop out more frequently than other students. In Berger's (1997) study of persistence, it was observed that non-White students are more likely to stay in their studies than White students.

When socio-economic and ability variables are taken into account, the retention picture is less clear (Hossler, 1984). Hilton (1982) and Ramist (1981) reported that when these variables are considered, dropout levels between African American and white students are about even. Johnson and Molnar (1996) compared retention rates for white, African American, and Hispanic students at Barry University in Florida, and discovered that when other variables were included in the analysis, race had little effect on retention. In summary, race/ethnicity was a central point in many retention studies and in majority of them certain races had higher persistence rates.

## Family Income Level/Socio-economic group

The research relating family income level to dropout has been fairly consistent. Astin (1975) found that as family income level decreased, dropout from higher education increased. Similar observation was also found by Feldman (1993) and Tinto (1975) who reported that students with lower family incomes are more likely to leave college than students from more affluent homes. In a national survey sponsored by the United States Department of Education's National Centre for Educational Statistics found that students in the highest socio-economic quartile drop out at a much lower rate than students from the lowest socio-economic quartile Porter (1990). Ramist (1981) also concurred with the above findings.

### Parents Educational Level

Pantages and Creedon (1978) reported that educational level of parents has little or no bearing on student persistence. Other researchers, including Astin (1975), Tinto (1975; 1993) and Ramist (1981) disagreed with the above findings. They argued that students with educated parents tend to value education more and are thus more likely to persist. Astin (1975) added, "*It seems likely that the more educated parents exert stronger pressure to stay in college than the less educated parents*" (pp. 35-36). Positive effect of parents education on student persistence is also observed by Elkins *et al.* (2000).

### Residence

Ramist (1981) reported that students who are from out states (that are not contiguous to the state in which the college is located) are more likely to dropout than the students from contiguous states. York (1993; 1999) also concluded his study with a similar findings as Ramist (1981). He also documented that students who leave outside of the campus location are more likely to leave early from the institution. Zheng *et al.* (2002) also found that staying in the university halls have a positive effect on student persistence.

### Accommodation

Past research shows that students who live in the university halls are more likely to integrate into the campus social system, therefore, more likely to persist than the students who live outside Tinto (1975; 1993). Berger (1997) found that residence halls are a source of social integration and it affects the process of social integration. He found that students who live in the university halls are more likely to persist. Elkins *et al.* (2000) also reported a similar findings as Berger (1997).

## Standardized Test Scores

The ACT (American College Testing) and SAT (Standardized Admission Test) tests are designed to predict academic success in college. Numerous studies have also linked scores on these tests to persistence in higher education. Ramist (1981) pointed out that the freshman year dropout rate of students scoring above 600 on the mathematical portion of the SAT is 9 percent, and the freshman year drop out rate of students scoring below 300 on the mathematical portion of the SAT test is 27 percent. Porter (1990) also found that student ability, as measured by a standardized test, effects student persistence. In the same study, the author concluded that, "*...low ability students have little likelihood of completing a degree in a timely manner and a high probability of dropping out*" (pp. 24). A clear and consistent linear relationship between SAT scores and college persistence was found by a study conducted by the Oregon State System of Higher Education (1994). Similarly, York (1993; 1999) concluded that as SAT scores increased so did the percentage of students earning degrees. At the University of Minnesota, DesJardins and Pontiff (1999) reported that students who score below the entering class average of 22 on the ACT test, account for a disproportionate number of dropouts. Thomas (2000) also reported a direct positive effect of SAT score on retention.

## High School Grade Point Average

Another measure associated with persistence in college is high school grade point average. Grade point average not only sheds high on a student's ability, but also the student's work ethic, and general attitude toward education. Thus, it is valued as an accurate predictor of success, as well as persistence in higher education. Numerous studies bear this out. Murtaugh *et al.* (1999) reported that persistence was found to increase with increasing high school grade point average. Astin (1993) conducted a longitudinal study of 39,243 students from over 100 institutions of higher education and discovered the following results: students with a C-high school GPA and an SAT composite of below 700 were five times more likely to drop out than students with an A average and SAT scores of 1300 or above. In a study of examining the effects of pre-enrolment variables at a community college, Feldman (1993) also found that the lower the high school GPA, the greater the chance the student would drop out. After tracking 1722 students for five years, York (1993; 1999) reported that high school GPA influenced degree outcome. A study conducted by the Oregon State System of Higher Education (1994) found a direct relationship between high school GPA and graduation rates. Similarly, Berger (1997) and Elkins *et al.* (2000) also found that high school GPA positively affect student persistence.

### High School Class Rank

Similar to high school GPA and standardized test scores, high school class rank has also been used as an admission factor in higher education. Like the standardized test scores and high school GPA, high school class rank has also been linked with persistence and attrition in higher education. DesJardins and Pontiff (1999) discovered that students who were in the bottom half of their graduating high school class dropped out disproportionately as compared to their counterparts who ranked in the top half of their high school graduating class. House (1994) also reported a significant correlation between high school class rank and persistence.

It is observed that when the pre-entry attributes/background variables (discussed above) are directly tested in Tinto's model, studies have indicated that these variables affect the level of initial commitment to the institution (Pascarella *et al.*, 1983; Terenzeni *et al.*, 1985; Braxton *et al.*, 1988) and directly affect student likelihood of persistence in college (Pascarella *et al.*, 1983; Nora *et al.*, 1990; Grosset, 1991; Caison, 2007).

### 3.7.2 Current Academic Experience

In terms of current academic experience two factors play an important role: current academic performance and extra curricular activities. In terms of current academic performance, Murtaugh *et al.* (1999) found that persistence increased with higher first-quarter grades. In a study at a comprehensive university, Roweton (1994) and Caison (Caison, 2007) discovered that college grade point average was the best predictor of retention of first year students. A similar observation was also found by Wall (1996) in a four-semester study with students at a community college. Wall (1996) documented that academic success as measured by the previous semester's grade point average is a strong determinant of retention. Defining academic integration in terms of grade point average and participation in an honor society, Whitaker (1987) found that it was one of the most influential variables related to persistence. In terms of extra curricular activities, Upcraft (1989) suggested, "*There is* *considerable evidence, however, that active participation in the extra curricular life of a campus can enhance retention and personal development*" (pp.150). Other scholars also agree with Upcraft (Tinto, 1993; Braxton *et al.*, 2000) and various studies bear this out (Berger, 1997).

### 3.7.3 Academic and Social Integration

Tinto hypothesizes that within the academic system, the student's academic performance (formal) and interaction with faculty/staff (informal) leads to either positive experiences that help to integrate the student into the intellectual community, or negative experiences that isolate the student. Similarly, within the social system, the student's involvement in formal extracurricular activities and informal peer-group interactions lead to positive experiences that lead to integration, or negative experiences that lead to disconnection. Numerous studies have tested the effect of academic and social integration in student retention.

When academic integration is directly tested in Tinto's model, studies have indicated that students who report a greater level of integration into the academic system of the institution will have a greater level of subsequent goal and institutional commitment (Pascarella *et al.*, 1983; Cabrera *et al.*, 1992a; Berger, 1997). In addition, it is expected that students who report a greater level of integration into the academic system of the institution will be more likely to persist in the institution (Cabrera *et al.*, 1992a; Berger, 1997; Thomas, 2000). Noel (1985) reported that a caring and helpful attitude expressed by faculty and staff is one of the most important retention tools on campus. In a follow-up phone interview with students who had left the institution, Li and Kilhan (1999) found that one of the key reasons for leaving was faculty did not care about their students. In developing a model to predict student retention for the Pennsylvania State System of Higher Education, Bailey *et al.* (1998) identified interaction with faculty and adequate advising as two major factors that contribute to student persistence. Similarly, Price (1993) concluded that close affiliation with faculty members related to student persistence.

Within the social system, the emphasis is on the student's involvement in informal peer-group interactions and formal extracurricular activities. Earlier research has confirmed the importance of social integration (Stage, 1988; Cabrera *et al.*, 1992b; Thomas, 2000). Astin (1993) and Ramist (1981) found that students having a greater level of integration into the social system of the institution will have a greater level of subsequent goal and institutional commitment and has been found to increase student persistence. Institutional and Goal Commitment

Institutional commitment refers to the student's commitment to the institution in which he or she is enrolled, and goal commitment refers to a student’s commitment to educational goals, such as the goal to graduate from college, or the goal to obtain a certain degree level. Tinto (1993) postulated that if the student experiences positive social and academic integration, institutional commitments and educational goals will be strengthened and the student will be more likely to stay enrolled. On the other hand, if the student's experiences in the academic and social systems are more negative, a student's goals and commitments will be weakened and the student will be less likely to remain at the institution (Tinto, 1993).

The concept of institutional commitment within Tinto's (1993) model has been particularly difficult to quantify. However, Terenzeni *et al.* (1981) have prompted out that knowledge of a student's institutional commitment provides one with information that further helps to identify leavers. In addition, these authors have shown that individuals who have a strong institutional commitment are more likely to graduate from that particular institution than those who have been identified with low or no institutional commitment (Terenzini *et al.*, 1981).

Along with institutional commitment, research has confirmed the importance of a student's goals in his or her persistence in higher education (Lenning *et al.*, 1980; Feldman, 1993; House, 1994). Waggener and Smith (1993) measured factors in retaining students at two important benchmarks during the academic year. They found that the most important factors associated with retaining students were the goal to obtain a degree, and a solid commitment to that goal at both of the benchmarks. Similarly, Zhang and Richard (1998) discovered that one of the key reasons for leaving from freshman year was the lack of personal commitment to a college education.

Using data from the Admissions Test Program Summary Report to analyse degree level goals, Ramist (1981) concluded that those who do not expect to obtain a four-year college degree are much more likely to drop out during their freshman year. In addition, Hossler (1984) reported that plans to enter graduate school are also related to persistence.

When institutional and goal commitments are directly tested in Tinto's model, studies have indicated that initial commitment to the goal of graduation from institution effects the level of academic integration (Pascarella *et al.*, 1983; Braxton *et al.*, 1988; Thomas, 2000), the level of social integration (Pascarella *et al.*, 1983; Thomas, 2000; Pascarella and Terenzini, 2005), and the subsequent level of commitment to the goal of graduation (Pascarella and Terenzini, 1983; Thomas, 2000; Pascarella and Terenzini, 2005). On the other hand, subsequent commitment to the goal of graduation from the institution increases the likelihood of persistence in the institution (Terenzeni *et al.*, 1985; Braxton *et al.*, 2000). In addition, many studies have confirmed that the subsequent commitment to the institution positively affects the level of academic integration (Braxton *et al.*, 1988; Allen and Nelson, 1989; Braxton *et al.*, 2000).

With respect to methodological perspective, it is evident that the majority of studies that tested Tinto’s model used a survey method to collect data. Though a survey design has been the primary data collection tool in these studies, it has some drawbacks, which is discussed in section 3.9. In addition, it is also observed that logistic regression and path analysis techniques have been used for data analysis in the most of the studies. Logistic regression analysis is specifically designed for use when the dependent variable has two values (Wright, 1995). Moreover, it is used due to their capability of handling both categorical and continuous variables. Also, it produces the predicted outcome as probabilistic values ranges from 0 to 1. Among other techniques, structural equation modelling, principal component analyses were used. Structural equation modelling requires a large sample size in order to get reliable and meaningful parameter estimation. At the same time, it is not always feasible for researchers to collect a large volume of dataset therefore, it is challenging to get reliable model with a small dataset using structural equation modelling (Brunsden *et al.*, 2000). Principal Component Analysis (PCA) is most useful when a large number of variables prohibit the effective interpretation of the relationships between objects. Researchers used this method to generate a smaller set of uncorrelated components rather than a large number of variables for the effective interpretation of the relationship among the variables.

## 3.8 Students' Marks/Performance Prediction

The topic of explanation and prediction of academic performance is widely researched. Increasing student success is a long-term goal in many educational institutions (Kovacic, 2010). Also, it is evident that current academic performance is the best predictor of predicting student success (Wall, 1996; Murtaugh *et al.*, 1999). If educational institutions can predict students' academic performance early, before their examination, then extra effort can be taken to arrange proper support for the lower performing students to improve their studies and help them to succeed. Therefore, the ability to predict student performance is very important in educational environments. Students' academic performance is based upon diverse factors like personal, social, psychological and other environmental variables.

Kotsiantis *et al.* (2004) used key demographic variables (age, gender, ethnicity etc.) and assignment marks to predict student's performance at the Hellenic Open University. In their study, they found using only the students' demographic variables, prediction accuracy varied from 58.84% to 64.47%. Using only background variables Al-Radaideh *et al.* (2006), Vandamme *et al.* (2007), Kovacic (2010), and Yadav and Pal (2012) also reported low prediction performance about 35%, 40.63% to 57.35% and 59.4% to 60.5% respectively in their study of performance prediction. Kotsiantis *et al.* (2004) observed that when other variables (for example, academic and environmental) beside demographics were included in model development, the prediction performance increased.

With this respect, Yadav and Pal (2012) conducted a study using 90 students data to predict students' academic performance. They used students' gender, admission type, previous school marks, medium of teaching, location of living, accommodation type, father's qualification, mother's qualification, father's occupation, mother's occupation, family annual income and student's family size. In their study, they achieved prediction accuracies 62.22% to 67.77%. However, Yadev *et al.* (2011) used students' attendance, class test marks, lab work, previous semester marks, seminar and assignment performance to predict students' performance at the end of the semester. In their study, they attained very low prediction accuracies from 45.83% to 56.25%. Bharadwaj and Pal (2011b) used 50 students data from Master of Computer Application department in VBS Purvanchal University, Jainpur from session 2007 to 2010. In their study, they used students' previous semester marks, class test grades, seminar performance, assignment performance, general proficiency, class attendance and lab work to predict students' mark in their end semester, and found a good prediction accuracy.

Past research evidenced that high school grades/previous education contributed the most in predicting students' performance among other variables (Al-Radaideh *et al.*, 2006; Vandamme *et al.*, 2007; Bharadwaj and Pal, 2011a). Vandamme *et al.* (2007) also found number of hours of mathematics, financial independence, and age of the first-year students in Belgian French-speaking universities were significantly related to academic success. In many studies parental income level was identified as the most significant factor to classify poor performing students (Quadri and Kalyankar (Quadri and Kalyankar, 2010; Bharadwaj and Pal, 2011a). Bharadwaj and Pal (2011a) also found living location, medium of teaching, mother's qualification, and student's family status were highly correlated with student academic performance. In a study on academic performance prediction conducted by Sembiring *et al.* (2011) found that interest, study behaviour, engagement time and family support are significantly correlated with student academic performance.

Researchers used different data analysis techniques in students' performance prediction studies such as Decision trees (DT) (Kotsiantis *et al.* (Kotsiantis *et al.*, 2004); Al-Radaideh *et al.* (Al-Radaideh *et al.*, 2006; Vandamme *et al.*, 2007; Quadri and Kalyankar, 2010; Yadav *et al.*, 2011; Yadev and Pal, 2012), Naïve Bayes (Kotsiantis *et al.*, 2004; Al-Radaideh *et al.*, 2006), Neural Networks (NN) (Kotsiantis *et al.*, 2004; Vandamme *et al.*, 2007), Support Vector Machine (SVM) (Kotsiantis *et al.*, 2004). The advantage of the NN for classification is their automatic training capacity, and ability to implement nonlinear decision functions (Looney, 1997; Jain and Duin, 2000). In recent years, SVM has been considered as a powerful tool for classification. The major strengths of SVM classification is that the training is relatively easy with few parameters and less possibility to get local optima. Unlike neural networks, SVM scales relatively well to high dimensional data and the trade-off between classifier complexity and error can be controlled explicitly (Gunn, 1998). To obtain best performance using SVM classifier, selection of a proper kernel is essential which is one of the major challenges of SVM classifier (Gunn, 1998). Among the classifiers Decision trees are very popular in the study of prediction performance because they produce classification rules that are easy to interpret (Yadev and Pal, 2012).

## 3.9 Issues with Traditional Survey-based Model

It is noted that, traditionally, retention studies are survey-based, where researchers use questionnaires to collect student data multiple times in a year and follow them for a specified period of time to determine whether they continue their study or not. Although the survey-based model has been successfully used to-date, survey-based research may be too burdensome to sustain, as individual institutions may not have the capacity to construct and administer a similar instrument to study their unique retention situation. Even if an institution is capable of fielding a one-time retention survey, repeated administrations over time may be too expensive and oppressive. In addition, survey-based models suffer with lacking of generalized applicability to other institutions. Moreover, another major limitation of survey-based study is low participation rates, which may often compromise the precision of the model output. Thus, it is key for enrollment professionals and researchers to have sufficient means of evaluating the trends in the circumstances of student retention at their institution in order to develop or adjust support programs accordingly.

## 3.10 Summary

To summarize, it can be said that student retention, progression and completion is one of the major issues in higher education institutions, which is being studied for many years. Researchers have developed many student retention models, but the most accepted and studied model is Tinto's retention model. While most research on Tinto's model is generally supportive, it should be noted that in every case the model leaves a great deal of explaining variance. According to Bean (1985), Tinto's model can usually explain no more than 0.35 of the variance. From the above reviewed literature, it can be noticed that most of the studies explained less than 35 percent of the variance while only a few studies have explained more than 35 percent of the variance. In relation to methodology, it can be noticed that most of the studies are survey-based and used Pascarella and Terenzini's scales (1980) to measure

Tinto's model constructs. Even though they have laid the foundation for the field, these survey-based research studies have been criticized for their lack of generalized applicability to other institutions and the difficulty and costliness of administering such large-scale survey instruments. An alternative approach to the traditional survey-based retention research is an analytic approach where the commonly available data is used.

Nowadays, institutions are collecting more data than ever. They routinely collect a broad range of information about their students, including demographics, educational background, social involvement, socioeconomic status, and academic progress. In addition, many external bodies are publishing open data on the web.

Although a large amount of data is available, data is frequently maintained in different locations, in different formats and with different identifiers. Integration of these data from different sources presents organizational challenges related to the ownership and use of the data. In recent years, linked data technologies

are considered to be well suited for data integration. In the next chapter, we will discuss linked data technologies, their suitability in higher education in general and their use in implementing predictive models in particular. In addition, we will discuss data analytics and their role in higher education institutions.

# Chapter 4: Linked Data, Open Data and Data Analytics

In recent years, linked data, open data and data analytics seem very promising in higher education and propose considerable research in this area. This chapter presents the concept and the rationale for linked data over HTML, spreadsheet and database, as these are the common data formats available in most institutional repositories or data sources. Furthermore, this chapter presents the rationale for linked data in higher education sector. In addition, it discusses open data and open data sources along with their potential in addressing student retention, progression and completion. Finally, this chapter discusses data analytics and the role of data analytics in higher education.

## 4.1 Semantic Web and Linked Data

The vision of Semantic Web is to contain structured data that could be analyzed and acted upon by software agents independently. It proposes extending the Web to include the data published using common formats based on Web principles (Berners-Lee *et al.*, 2001; Shadbolt *et al.*, 2006). Berners-Lee *et al.* (2001) summarised the vision of Sematic Web as *"... an extension of the current Web in which information is given well-defined meaning, better enabling computer and people to work in cooperation.*"

The Semantic Web vision advocates publishing structured information on the Web based on adopting a common stack of technologies (Figure 4.1). The core principles include representing structured information with a common data model, resources being uniquely identified with global, Web identifiers, the ability of these resources to refer to one another and describing the conceptual characteristics in commonly accepted languages. In effect, the Semantic Web promotes establishing a global distributed database of structured information sources to be globally accessible to applications. Semantic Web proposal envisions a machine-readable layer over which applications can share and

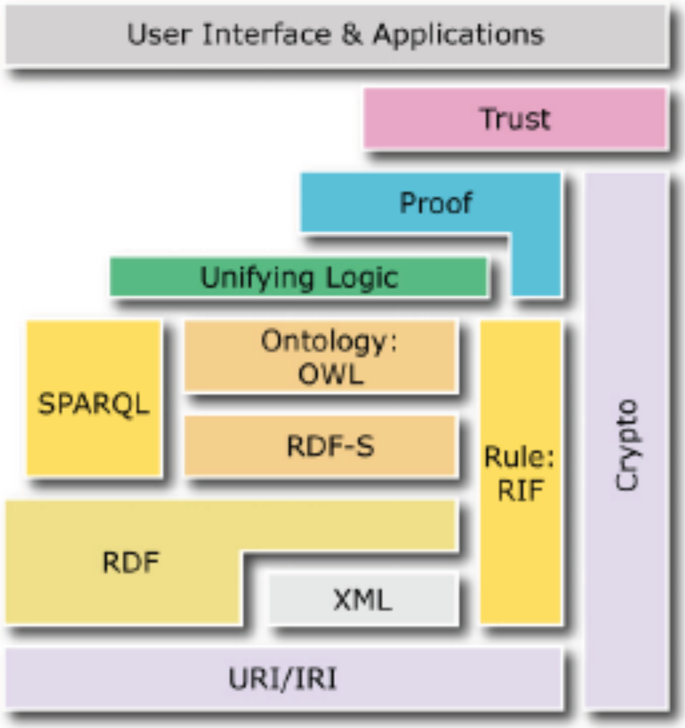


Figure 4.1: The Semantic Web Technology layer cake.

reuse data. The basic foundations of the Semantic Web, however, start with publishing structured data using common standards and in a way that is compliant with the general architecture on the Web. Because of the many technical limitations of efficiently utilising the upper stack technologies and still evolving standards, the Linked Data initiative aims to utilise the lower stack of technologies as a way of publishing for the purposes of basic data integration and data reusing. Berners-Lee (2009) refers Linked Data (LD) to the best practice of publishing structured data on the Web and linking them together to obtain new knowledge from different data sources. Often the terms Linked Data and Semantic Web are used interchangeably, however strictly speaking Linked Data is just a way of publishing data on the Web according to the following principles outlined by Berners-Lee (2009):

- Use URIs as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL).
- Include links to other URIs, so that they can discover more things.

Therefore, it can be said that semantic Web is a vision of creating a Web of Data and Its ultimate goal is to automate the Web-scale processing and integration of information. On the other hand, Linked Data is a concrete means to achieve that vision by proposing to share a common methodology for publishing data that avoids the heterogeneity of data sources[26].

According to Bizer *et al.* (2009), linked data refers to data published on the Web in a manner that it is machine readable, its meaning is explicitly defined, it is linked to other external data sets, and can also be linked from external data sets. Exposing data as Resource Description Framework (RDF) is an important first step, but to actually achieve the linked-data vision it is required to set explicit RDF links between data items within different data sources. This provides the means by which more information can be discovered about a given entity. Each unit of Linked Data expressed in RDF has a subject, predicate, and object. All subjects, predicates, and objects (other than simple data values) are encoded or represented as uniform resource identifiers, or URIs, intended to be resolvable as uniform resource locators (URLs). To understand the meaning of these data ontology plays a centre role. Gruber (1993) defined the term ontology as a specification of a conceptualization. Ontologies are used for the explicit description of the information source semantics. Ontologies capture essential information including what type of data is contained, what are the relationships between entities in the data, and any specific rules (inference rules) to conduct automated reasoning (Berners-Lee *et al.*, 2001). For example, an ontology may express the rule that 'if a post code is associated with a participation neighbourhood group code, and a student’s permanent address includes that post code, then that student has the associated participation neighbourhood group code'. The receiving application can then infer that, if a particular postcode is provided, that student must be from a particular participation neighbourhood group. Effectively, all that ontologies allow an application to do is manipulate the information provided according to predetermined rules and come to a logical conclusion about that data in the format that it requires. In addition to breaking down silos, linked data has also through its fundamental dependence on ontologies charted new ground in practices for data

[26] http://vadimeisenberg.blogspot.ca/2011/10/on-difference-between-linked-data-and.html

description or metadata. Changes to metadata practice driven by the adoption of linked data can best be summarized as making once implicit statements explicit. Declaring the subject of every metadata statement with a URI as its identifier and using defined types and properties (also specified by URIs) for expressing the content of metadata in RDF eliminates much of the ambiguity in what is being referred to, in where the intended meaning has been defined, and in how the information referenced can be directly accessed.

The basic assumption behind linked data is that the value and usefulness of data increases the more it is interlinked with other data (Bizer *et al.*, 2007). This principles necessitates a common data format based on URIs and RDF (World Wide Web Consortium, 2004), as well as use of SPARQL (World Wide Web Consortium, 2008) as a common language to manipulate the data. In addition, the fourth principle encourages data providers to federate/join their datasets to others in the Web of Data by explicitly stating the relationships between the data they publish and the data already published by others. To accomplish this principle, ontologies play a central role. Therefore, datasets in the Web of Data are federated, and a data consumer application can automatically discover, access and integrate data from other sources (Heath and Bizer, 2011).

In order to enable web-scale data federation, the W3C (World Wide Web Consortium) proposed several standards that homogenize the data formats and the data access on the web. The most relevant ones for this thesis dissertation are now explained:

- Uniform Resource Identifier (URI): A URI is a compact string of characters for identifying an abstract or physical resource. From the very beginning of the Web, URIs have played a crucial role because they were the way of interlinking the available documents and resources. The Semantic Web extends this concept of resource to whatever might be identified by a URI, including people, media, companies, relationships, actions and any other concepts that can be identified by a URI: it does not matter if it is accessible via Web or not. Thus, URIs provide a common way to univocally identify any resource in the Web.

- Resource Description Framework (RDF): RDF is a simple data model to publish structured data on the Web. The underlying structure of RDF is based on triples, each consisting of three elements: a subject, a predicate and an object triples. The subject of a triple is the URI identifying the described resource. The object can either be a simple literal value, such as a string, number, or date, or the URI of another resource that is somehow related to the subject. The predicate indicates what kind of relation exists between subject and object, e.g. this is the name or date of birth (in the case of a literal), or someone the person knows (in the case of another resource). The predicate is a URI as well. Thus, RDF facilitates expression of a simple fact in a flexible way. RDF can be represented using a number of languages such as, RDF/XML, turtle, N3.

- SPARQL Query Language: SPARQL is a query language for RDF datasets that is now widely used. It defines a syntax and semantic to query RDF data sources and to process the obtained results. SPARQL queries can be keyword-based, or they can restrict results depending on their relationships to other concepts. Additionally, SPARQL facilitates querying sets of triples (called “graphs”), as well as constructing new triples out of the ones retrieved.

## 4.2 The Rationale for Linked Data

Over the last ten years there has been a growing realisation regarding linked data’s power for exposing, sharing, and connecting pieces of data and information using uniform resource identifiers (URIs). In order to understand the concept and value of linked data, it is important to consider contemporary mechanisms for sharing and reusing data. Heath and Bizer (2011) have stated the problem with re-using data published in HTML format. They stated that a key factor in the re-usability of data is the extent to which it is well structured.

- The HTML website is concerned with structuring textual documents rather than data. As data is amalgamated into the surrounding text, it is hard for software applications to extract snippets of structured data from HTML pages. To address this issue, a variety of *microformats*[27] have been invented. Limitations of microformats are that they are restricted to representing data about a small set of different types of entities; they only provide a small set of attributes that may be used to describe these entities; and that it is often not possible to express relationships between entities, for example, a student can be an employee of a university, rather than being just a student of the university. Therefore, microformats are not suitable for sharing arbitrary data. Moreover, the new specifications of HTML5 implement a local storage feature that is a key-value pair storage on the browser. Despite these advances, it is still difficult to get raw data out of the information (Huynh *et al.*, 2007). Schema.org[28] is another advancement, which stores vocabularies and its goal is to improve the display of search results, making it easier for people to find the right web pages. However, schema.org has limitation on always using distinguishing URLs for things from URLs for pages about those things.  Also, schema.org is the focus on a single integrated core vocabulary, rather than an overlapping patchwork of independent schemas.

- In HTML website link refers to the document of the data and human interference needed to find out the actual data from that document. In other way we can say, HTML links typically indicate that two documents are related in some way, but mostly leave the user to infer the nature of the relationship.

According to Omitola *et al.* (2010) data published in spreadsheet format also have inherent problems with respect to re-usage, including:

- little or no explicit semantic description, or schema, of the data. An example of this can be when IDs are labelled such as “Std ID” or “Prog ID” without definitions and no explanation of the relationship with the rest of the data in the spreadsheet.

- more difficult to integrate, or link, data from disparate data sources. An example of this can be where employment status value for each student was given as “Employed” or “Non-Employed”. It would be good to know how this data was arrived,

[27] http://microformats.org/

[28] http://schema.org/

and linking it with the data sources from whence they come would have been useful (e.g. for provenance and validation).

Linked Data provides the following solutions to realize the above problems:

- RDF provides a flexible way to describe things in the world – such as people, locations, or abstract concepts and how they relate to other things. The key features of RDF worth noting in this context are the following:
    - RDF links things, not just documents: RDF links do not simply connect the data fragments, but assert connections between the entities described in the data fragments (Heath and Bizer, 2011).
    - RDF creates typed links between data from different sources (Bizer *et al.*, 2009; Heath and Bizer, 2011). These may be as diverse as databases maintained by two organizations in different geographical locations, or simply heterogeneous systems within one organization that, historically, have not easily interoperated at the data level.
- Taxonomies, vocabularies and ontologies provide domain-specific terms for describing classes of things in RDF and how they relate to each other in *SKOS*[29] (Simple Knowledge Organization System), *RDFS*[30] (the RDF Vocabulary Description Language, also known as RDF Schema) and *OWL*[31] (the Web Ontology Language). According to Berners-Lee *et al.* (2001), an ontology refers to a document or file that *formally defines the relations among the data*. The typical Web ontology consists of both a taxonomy and a set of inference rules that computers can use to conduct automated reasoning to create new knowledge from the existing information. On the other hand, the taxonomy defines all the classes of objects and any

[29] http://www.w3.org/TR/skos-reference/

[30] http://www.w3.org/TR/rdf-schema/

[31] http://www.w3.org/TR/2004/REC-owl-features-20040210/

> relationships between them (Berners-Lee *et al.*, 2001). The inference rules allow an application to make decisions based on the classes supplied without needing to actually understand any of the information provided (Berners-Lee *et al.*, 2001). There is no clear division between what is referred to as "ontologies" and "vocabularies". Vocabularies also define the concepts and relationships used to describe and represent an area of concern. The trend is to use the word "ontology" for more complex, and possibly quite formal collection of concepts and relationships, whereas "vocabulary" is used when such strict formalism is not necessarily used or only in a very loose sense. The Web Ontology Language, OWL provides greater expressivity of data compared to the Vocabulary Description Language, RDFS.

Also, discussing the advantages of RDF model over relational database model Tim Berners-Lee (1998) stated a difference between XML/RDF schemas and Relational Database (RDB) schemas. He stated that many web sites can export documents structured by the same schema. On the other hand, a database schema is created independently for each database. Adopting the example from him, if a million companies clone the same form of employee database, there will be a million schemas, one for each database. It may be that RDF will fill a simple role in simply expressing the equivalence of the terms in each database schema. Therefore, in the case of RDF there will be a relatively small number of XML/RDF schemas compared to database and this in fact provides better interoperability as they have same schema. There is considerable interest in RDB2RDF conversion at present as RDB is the most common format for data storage in many organizations, and the W3C has recently set up an Incubator Group to work on it. The Incubator group reported current techniques, tools and applications for mapping between RDB and RDF in (Satya *et al.*, 2009).

## 4.3 The Rationale for Linked Data in Higher Education

It can be said that linked data provides more expressivity of data and enables typed links to be set between items in different data sources, and therefore connect these sources into a single global data space. The use of Web standards and a common data model (RDF) make it possible to implement generic applications that operate over the complete data space, and this is the value of Linked data.

In the higher education sector, the understanding of linked data and its implications is not currently widespread, as stated by Miller (2010) in the report, "*Linked Data Horizon Scan*". In the report, Miller also stated many opportunities of linked data in higher education. The research paper stored in institutional repository would be linked to related papers by the same authors, and placed within context to demonstrate institutional research ability. The courses offered by one institution would be automatically aggregated with similar courses from elsewhere and made easily accessible to potential students. Relevant data from institution would be available alongside that from other bodies, powering a range of applications for staff, students, funders, industrial partners and more.

Also, Tiropanis *et al*. (2009a; 2009b; 2009c; 2009d) suggested that building a field of linked open data across UK HE/FE institutions by selectively and securely exposing repositories and institutional data can provide significant value and pave the way for pedagogically meaningful applications powered by application-wide or community-wide agreed ontologies in the future. HE/FE challenges can be addressed by efficiently linking information across institutions. Learners and teachers will be able to efficiently search across various repositories. Learning and teaching will be better supported with utilities that enable targeted searching on authoritative teaching and learning material across institutional repositories. Prospective students and module designers will be able to make comparisons of curricula if such information is exposed in linked data formats.

More recently, the JISC CETIS news, "*Big Data and Analytics in Education and Learning*[32]" in 2011 reported that the characteristics (volume, variety and velocity) of Big data[33] require new methods and infrastructure, new data management tools, and new skills to manage and analyze data. In the same report, it was stated that in order to gain the opportunity of the Big data in HEI,

[32] http://blogs.cetis.ac.uk/cetisli/2011/12/14/big-data-and-analytics-in-education-and-learning/

[33] Big data describes the datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.

HEI need to employ new approaches, new tools, and new skills. Linked data denoted as one of the technologies to meet the Big data's requirements. Linked data technologies can help in linking disparate data sources to enhance the interoperability and discovering more data in higher education relating to student success. Linked data can help to identify at-risk students from the very beginning by linking student specific data from different data sources. For example, HE institutions can monitor students who are obtaining poor marks in their assignment by querying over student records and also they might find a possible reason why these students obtained poor marks by querying over other related linked data sources. Moreover, higher education institutions can develop student prediction models to early identify students who are likely to stay/dropout from their study by integrating different internal or external institutional data sources so that they can provide more support to those students to succeed in their study.

McAuley *et al.* (2011) also reported about the existing opportunities of open data and linked data in higher education. They reported that, substantial learning challenges could be met by interlinking resources across disciplines and institutions.

Most recently, William Hammonds also reported in "Open data in higher education – the 'next big thing'?[34]" the value of open data and linked data in higher education sector. He reported that the importance of using data effectively is well known to universities, universities know the power of data and the benefits of exposing data to analysis and reuse to drive discovery, innovation and advance knowledge. He noted that Open data allows a wider range of users to analyze it and potentially generate new uses, while linked data methods allow different data sets to be combined to further expand these potential benefits.

## 4.4 Open Data and Open Data Sources

McAuley *et al.* (2011) referred to Open Data (OD) as *"the philosophical and methodological approach to democratizing data, enabling individuals,*

[34] http://blog.universitiesuk.ac.uk/2013/07/17/open-data-in-higher-education/

*communities and organisations to access and create value through the reuse of non-sensitive, publicly available information"*.

Open data is typically available online at no cost to citizen groups, non-governmental organisations (NGOs) and businesses. It is expected that open data will support greater transparency and accountability of data, and provide reusability of data. To date, the open data movement has created great excitement in developer communities. Social and commercial entrepreneurs are producing a seemingly endless stream of innovative applications that repurpose and enrich publicly available data, across multiple sectors, including health, transport, education and the environment. Higher education has pioneered the use of web technologies, with institutions making large amounts of information available to students, commercial partners, funding agencies and staff.

In the UK, a number of external bodies regularly publish open data in the web such as ***Higher Education Funding Council for England (HEFCE[35]),* which** publishes students' participation group in a five quintile ordered from '1' (those with the lowest participation) to '5' (those with the highest participation) based on their postcode; ***the Office for National Statistics (ONS***[36]), which publishes annual income by profession or occupation; ***Unistats***[37], which publishes National Student Survey (NSS) result that is conducted to measure students' satisfaction in teaching and learning, assessment and feedback, academic support, organization and management, learning resources and personal development; **data.gov.uk**[38], which publishes public sector datasets, ***Higher Education Statistics Agency (HESA***[39]), which publishes a quantitative information about UK higher education such as students' non-continuation rate by age marker, previous HE marker, low participation marker, entry qualification and subject of study. This information could be used in addressing a number of higher education challenges such as student retention,

[35] http://www.hefce.ac.uk/whatwedo/wp/ourresearch/polar/polar2/

[36]http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-250731

[37] http://unistats.direct.gov.uk/

[38] data.gov.uk

[39]http://www.hesa.ac.uk/index.php?option=com_content&task=view&id=2064&Itemid=141

progression and completion through integrating these data to other datasets and analyzing the new set of data.

## 4.5 Open Data vs Linked Open Data

Linked Open Data (LOD) is the data published on the Web in accordance with the linked data principles under an open licence, which does not hinder its reuse for free (Berners-Lee, 2009). In other words, it can be denoted as if the linked data principles are applied to open data then it will be called linked open data. The most common opinion is that RDF is a standard for representing linked data.

To fully benefit from open data, it is crucial to put information and data into a context that creates new knowledge and enables powerful services and applications. As linked open data facilitates innovation and knowledge creation from interlinked data, it is an important mechanism for information management and integration.

Sir Tim Berners-Lee best described the path from open data to linked open data in his first presentation of 5 Stars Model at the Gov 2.0 Expo in Washington DC in 2010. Since then, Berners-Lee‘s model has been adapted and explained in several ways; the following adaptation of the five Stars Model[40] by Michael Hausenblas explains the costs and benefits for both publishers and consumers of linked open data.

Universities know the power of data through their research activities and the benefits of exposing data to analysis and reuse to drive discovery, innovation and advance knowledge. In the UK, there are few universities currently exposing their public data as linked data, using technologies such as RDF and SPARQL to give direct access to the information. For example, the Open University[41], the University of Southampton[42] and the University of Oxford[43] have looked at how open and linked data practice can be applied to their own institutions’ administrative data.

[40] http://5stardata.info/
[41] http://data.open.ac.uk/
[42] http://data.southampton.ac.uk/
[43] http://data.ox.ac.uk

| | |
|---|---|
| ★ | Information is available on the web (whatever format) under an open license. |
| ★★ | Information is available as machine-readable structured format (e.g., Excel instead of image scan of a table). |
| ★★★ | Non-proprietary formats are used (e.g., CSV instead of Excel). |
| ★★★★ | URIs are used to denote the things, so that people can point at individual data. |
| ★★★★★ | Data is linked to other data to provide context. |

Figure 4.1 Sir Tim Berners-Lee's Five Stars Model.

Outside the UK, several other universities and education institutions are also publishing their information with linked data (RDF, SPARQL), such as Linked Open Data at University of Muenster[44] and the LODUM[45] project in Germany or the Norwegian University of Science and Technology[46] exposing its library data as linked open data. Furthermore, educational resources metadata has been exposed by the mEducator project (Mitsopoulou *et al.*, 2011; Dietze *et al.*, 2012) (Mitsopoulou, *et al.*, 2011; Dietze *et al.* 2012). A more thorough overview of educational Linked Data is offered by the Linked Education[47] platform.

These initiatives are currently often disconnected from each other. The potential for linked data in education and research goes well beyond the individual benefit for each institution, as this potential can only be achieved through providing cross-university data that can be aggregated, integrated and compared. The data.ac.uk[48] initiative is developing into a central point for open and linked data sets, which encourages community to share, utilise, update, grow and generate demand for open data. JISC and initiatives such as Linked

[44] http://data.uni-muenster.de
[45] http://lodum.de
[46] http://openbiblio.net/2011/09/08/ntnu/
[47] http://linkededucation.org
[48] http://www.data.ac.uk/

Universities[49] and the LinkedUp Project[50] are also developing and disseminating good practice in this respect.

## 4.6 Data Analytics and the Role of Data Analytics in Higher Education

The data analytics simply means the use of analysis, data, and systemic reasoning to make a decision (Campbell and Oblinger, 2007; Davenport *et al.*, 2010). It is an overarching concept that van Barneveld *et al.,* (2012) have defined simply as "data-driven decision making (DDD)". However, Cooper (2012) insists that analytics is not just about making decisions; it is inclusive of exploration and problem identification. JISC CETIS's deputy director, Adam cooper (2012) defined analytics as *"the process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/or simulated future data".* Moreover, Campbell *et al.,* (2007) stated "*analytics marries large datasets, statistical techniques and predictive modelling. It could be thought as the practice of mining institutional data to produce actionable intelligence* ".

With the current deluge of data from disparate sources, analytics have a promising potential to increase the value of such data. For instance, analytics play a role in facilitating the examining of decisions before they are made, which might help in making smart decisions. The data-driven decision making (DDD) might be better than the experience-based decision for many reasons (Davenport *et al.*, 2010). Indeed, research has demonstrated that organizations that make their decisions based on analysis of the data, have shown higher performance than the organizations that don't. In other word, DDD plays a key role in increasing the output and the productivity of organizations (Brynjolfsson *et al.*, 2011). According to Davenport *et al.* (2010), six questions can be answered through the effective use of analytics organized by time-frame and by information vs. insight (see Table 4.1), and answering these questions may help many organisations to address many of their problems.

[49] http://linkeduniversities.org/lu/

[50] http://linkedup-project.eu/

Table 4.1 Six key questions can be addressed by analytics (Davenport *et al.*, 2010. Reproduced with permission).

| | **Past** | **Present** | **Future** |
|---|---|---|---|
| **Information** | What happened?<br>(Reporting) | What is happening now?<br>(Alert) | What will happen?<br>(Extrapolation) |
| **Insight** | How and why did it happen?<br>(Modeling, Experimental design) | What is the next best action?<br>(Recommendation) | What is the best/worst that can happen?<br>(Prediction, Optimization, Simulation) |

The use of analytics in higher education is a relatively new area of practice and research. Higher education institutions are adopting the practices to ensure organizational success at all levels by addressing questions about student retention, admissions, funding and operational efficiency (van Barneveld *et al.*, 2012). In an age, where educational institutions are under growing pressure to reduce costs and increase efficiency, Natsu (2010) reported that analytics can help education leaders to reduce costs and improve teaching and learning. She stated that analytics can help in enhancing student achievement, planning courses, recruiting and retaining students, optimizing the scheduling of classrooms, and maximizing alumni donations.

Analytics in the education domain is providing increased opportunities for learning and teaching, and offers more convenient evidence-based decision-making, action and personalisation in different areas of education. In the education domain, generally two types of analytics are used (van Harmelen and Workman, 2012): Learning analytics (LA) and Academic analytics (AA).

LA is the application of analytic techniques to analyse data about learner and teacher activities, to identify patterns of behaviour and provide actionable information to improve learning and learning-related activities. The first International Conference on Learning Analytics and Knowledge (2011) defined learning analytics as "*the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs*". The main goal of LA is to identify the learners who are at-risk in their programme of study as early as possible to allow for implementing some early intervention strategies, which may help students to succeed and retain in their study (Johnson *et al.*, 2011; van Barneveld *et al.*, 2012).

Table 4.2 Differentiate between learning and academic analytics (Long and Siemens, 2011. Reproduced with permission).

| Types of Analytics | Level or Object of Analysis | Who Benefits? |
|---|---|---|
| Learning Analytics | *Course-level:* social networks, conceptual development, discourse analysis, "intelligent curriculum" | Learners, faculty |
| | *Departmental:* predictive modelling, patterns of success/failure | Learners, faculty |
| Academic Analytics | *Institutional:* learner profiles, performance of academics, knowledge flow | Administrators, funders, marketing |
| | *Regional (state/provincial):* comparisons between systems | Funders, administrators |
| | *National and International* | National governments, education authorities |

Whereas, LA is primarily concerned with increasing learner success and the achievement of specific learning goals (van Barneveld *et al.*, 2012), AA's aim is corresponding to that of business analytics in the corporate sector: increasing organizational effectiveness (Long and Siemens, 2011). AA focuses the role of data analytics at the institutional, administrative and policy-making levels. AA helps higher education institutions to fulfil their mission in different area of higher education such as, student recruitment, student retention and budgeting (van Harmelen and Workman, 2012). Table 4.2 suggested by Long and Siemens (2011) clarifies the difference between academic and learning analytics.

The importance of analytics for the success of higher education is growing. Drawing value from data in order to guide planning, interventions, and decision-making is an important and fundamental shift in how education systems function. According to Bichsel (2012), there are many note-worthy examples of successful analytics use across a diverse range of institutions. For example, Paul Smith's College used analytics to improve its early-alert program providing more efficient and more effective interventions that resulted in increased success, persistence, and graduation rates; the university of

Washington Tacoma and Persistence Plus partnered to improve persistence and grades in online math courses through notifications on user-identified personal devices; the Open University used analytics to identify at-risk students in virtual learning environments using existing data. In addition, the Signals project (Arnold, 2010) at Purdue University used analytics in identifying at-risk students to improve student success, retention and graduation rates.

Growing interest in data and analytics in education, teaching, and learning raises the priority for increased, high-quality research into the models, methods, technologies, and impact of analytics. Two distinct research communities, Educational Data Mining (EDM[51]) and Learning Analytics and Knowledge (LAK[52]), have developed in response. EDM and LAK both reflect the development of data-intensive approaches to education. LAK and EDM share the goals of improving education by improving assessment, how problems in education are understood, and how interventions are planned and selected. An increased volume of data sets available from students' interactions with educational software and online learning and from public data repositories raises the need for research-based models and strategies. Both communities have the goal of improving the quality of analysis of large-scale educational data, to support both basic research and practice in education (Siemens and Baker, 2012).

## 4.7 Summary

This chapter discusses linked data and the prospects of linked data in higher education. It also focuses on open data and available external open data sources/repositories that could be used in supporting higher education challenges such as student retention, progression and completion. The value of these open data sources to higher education lies not merely in openness and accessibility, but in their interconnectivity. The capability to query as well as browse, to benefit from data fusion mechanisms, generates both novel research discoveries and compelling educational experience. Furthermore, this chapter describes analytics and the role of analytics in higher education.

[51] http://www.educationaldatamining.org/

[52] http://www.solaresearch.org/events/lak/

# Chapter 5: Identifying Institutional Repositories and External Open Data Sources to address the Higher Education Challenges

This chapter provides a comprehensive summary of the data and data repositories or data sources, which can be used in addressing higher education challenges and present them in a structured way. This chapter also discusses the opportunities and challenges in sharing repositories and reports approaches to address these challenges of sharing repositories.

Section 5.1 explains the methodology to identify and classify the repositories/data sources to address the higher education challenges. Section 5.2 specifies institutional data and data repositories, which is being used or can be used in addressing the higher education challenges based on the literature; while section 5.3 states the external open data and open data sources, which can be used to address the higher education challenges specifically to address student retention, progression and completion. Finally, section 5.4 reports the opportunities and difficulties of sharing repositories/data sources. The section also discusses the approaches to address the challenges of sharing repositories.

## 5.1 Methodology

To identify the data and data sources that are required to address the higher education challenges, we have reviewed a range of educational literature, which are related to higher education challenges. At the first stage, we identify the data, which is being used or can be used to address the HE challenges and grouped them into two broad categories based on their possible sources: a) Institutional repositories: datasets that are stored in the institutional internal data sources and produced by higher education institutions and b) External repositories/open data sources: datasets, which are freely available on the web to use and published by external bodies other than higher education institutions. Furthermore, we categorised institutional repositories into 9 key repositories based on their institutional data sources and relativity of the data in addressing higher education challenges.

## 5.2 Institutional Repositories

According to McCord (2003), *"an institutional repository is a digital archive of the intellectual product created by the faculty, research staff, and students of an institution and accessible to end-users both within and outside of the institution with few if any barriers to access".* Specifically, in this thesis the term "institutional repository" is used to refer the available data sources in the higher education institutions. The institutional data and data repositories, which can be used in addressing higher education challenges, are grouped into 9 key repositories, which are course information, teaching and learning material, student records, research information, virtual learning environments, accreditation records, academic staff, research staff and expertise information, Staff facilities and development programme and resource information. It is noted that some of these institutional repositories are currently defined as institutional data sources in the higher education institutions. It is noteworthy that the sense in which the term "institutional repository" is used in this thesis is not widely shared "nowadays". However, in this thesis the term *institutional repositories* and *institutional data sources* are used interchangeably as both of them are commonly used to store institutional data. Figure 5.1 displays the 9 key institutional repositories, whereas Table 5.1 presents the summary of which institutional repository can be used to address which higher education challenges. The description of these repositories is illustrated below:

### 5.2.1 Course Information

Course information institutional repository mostly contains courses/programmes information. This repository also includes the goals of the programme, the intended learning outcomes, syllabi, learning and teaching methods, types of assessment, time tables, programme fees and length of the programmes.

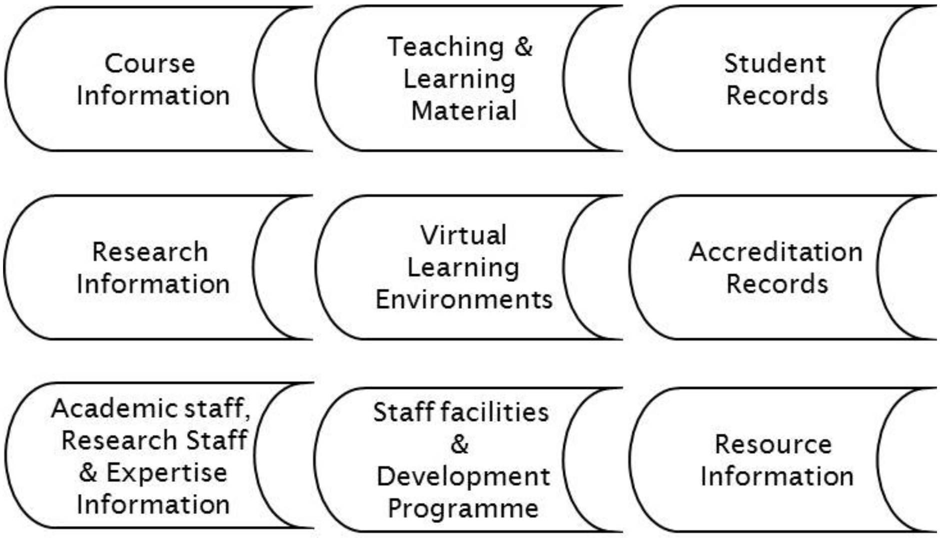

Course Information
Teaching & Learning Material
Student Records
Research Information
Virtual Learning Environments
Accreditation Records
Academic staff, Research Staff & Expertise Information
Staff facilities & Development Programme
Resource Information

Figure 5.1 Institutional repositories to address the higher education challenges.

This type of repository answers some of the higher education challenges. Specifically, according to BIS (2009), Tiropanis *et al*. (2009d) and HEFCE (2009b), course information available in different institutions can be used to design more efficient curriculum, programme or module. The module designers can compare programmes or modules in different institutions and find the gap and can offer new programmes or modules (Tiropanis *et al*., 2009a; 2009d). Moreover, to attract local/international students institutions need to make courses/programmes information accessible to everyone (Hirsch and Weber, 1999; BIS, 2009; Tiropanis *et al*., 2009d). Courses/programmes information also needs to be made accessible to employers to enhance the student employability (Hirsch and Weber, 1999; Rae, 2007; BIS, 2009; HEFCE, 2009b; 2010b). Furthermore, this information needs to be made accessible to the accreditation bodies for more efficient accreditation of the higher education institutions and programmes (ACE *et al*., 2006; Guerra-López, 2008; Eaton, 2009). According to HEFCE, the programmes output that deliver the higher-level skills need to be made accessible to funding bodies to attract funding.

## 5.2.2 Teaching and Learning Material

These types of repositories contain teaching and learning material of an institution. According to (Hirsch and Weber, 1999; Hanna, 2003; Biggs and Tang, 2007; HEFCE, 2009a; 2009b; Tiropanis *et al*., 2009d) enabling access to teaching and learning materials across institutions will certainly improve the quality of learning and teaching of the higher education institutions as students and teachers can access a large amount of learning and teaching material available across the institutions and they can develop themselves accordingly. West (1999) and Gumport and Chun (2005) also believe that teaching and learning materials need to be shared across the institutions for better quality of learning and teaching activities in the institutions as teachers and learners can have more deeper understanding on any specific subjects. They can broaden their knowledge having lots of information on any subject area.

## 5.2.3 Student Records

### 5.2.3.1 Student admission data

Repositories that contain students' general information for example, personal data (name, contacts, email, homepage, URL, images), relationships to other people in the institutions, interests, accessibility and preferences (language preference, disability, eligibility), demographic characteristics (e.g., ethnicity/race, sex, age), geographic origin/residency, financial information, students' living arrangements, students security feature (keys, password, credentials) etc. As per the American Council on Education (ACE) *et al.* (2006) and Ounnas *et al.* (2006), the repository containing student admission data needs to be made accessible across departments (for students with common modules) in the institutions to efficiently create groups for learning and teaching activities. For example, if the teacher wants to build groups according to students' geographic origin, then this repository will help them to create the groups efficiently and so on. Also, to efficiently support student retention, progression and completion, institutions need to make student admission data accessible, so that institutions can analyse different data to monitor students progress (Murtaugh *et al.*, 1999; Thomas, 2002; Hanna, 2003; Crosling *et al.*, 2009; HEFCE, 2009b). For example, according to Tiropanis *et al.* (2009d), institutions can analyse student interest and other information to identify the early sign of student disengagement with their study. Student admission data and contextual information of students (like academic attainment, aptitude and potential) also need to be considered in widening participation in higher education (Education and Skills Committee, 2003; BIS, 2009).

#### 5.2.3.2 Student academic record

Repositories contain students' academic information including goal of the learner, achievements and learner history performance (students' pre-college characteristics/academic preparedness (e.g. high school GPA, SAT score)), certifications, competency/skills/experience/knowledge, portfolio, current programme's information, transcript (grades), activities, involvement in campus programs (e.g., freshman orientation course, educational opportunities pro- gram), context, and extra-curricular activities etc. As per American Council on Education (ACE) *et al.* (2006) and Ounnas *et al.* (2006), institutions need to make this information assessable across departments to efficiently create group for learning and teaching activities. For example, if the teacher wants to build a student group according to students GPA (e.g. who have high GPA or any order) then this repository will help them to create the group efficiently. This repository also needs to support student retention, progression and completion effectively and efficiently. According to QAA and HEFCE, students' academic information are needed to efficiently support in student retention as institutions can monitor students' progress on any subject from their grades. If any student possesses poor grades, they can find out the reason and arrange support accordingly. Also, some selective information from this repository is needed to be available to the employers to enhance student employability (skills, knowledge, work experience and personal attributes) (Hirsch and Weber, 1999; West, 1999; Bridges, 2000; Biggs and Tang, 2007; HEFCE, 2009a; 2009b).

### 5.2.4 Virtual Learning Environments

Virtual learning environments (VLE) are widespread in HEI for supporting and facilitating both teaching and learning (Ho *et al.*, 2009). According to What is.com[53], "*a virtual learning environment is a set of teaching and learning tools designed to enhance a student's learning experience by including computers and the Internet in the learning process. The principal components of a VLE package include curriculum mapping (breaking curriculum into sections that can be assigned and assessed), student tracking, online support for both teacher and student, electronic communication (e-mail, threaded discussions, chat, Web publishing), and Internet links to outside curriculum resources*".

[53] http://whatis.techtarget.com/definition/0,,sid9_gci866691,00.html

VLE offers a range of learning tools and facilities that aid in delivering, communicating, and managing the course (Britain and Liber, 1999; O'Leary, 2002; JISC infoNet, 2004). For instance, communication facilities between students and tutors, between students and students or across student groups through e-mail, discussion board and virtual chat; announcements and a noticeboard facility; assessments and testing facilities through multiple-choice assessment with automated marking and immediate feedback; scheduling/calendar; assignment submission; class list and student homepages facility allow students to know other students in the same course or tutors to have some idea about students' backgrounds, interests and aspirations; integrated web 2.0 tools such as wikis, blogs, whiteboard, and authoring tools and many other features. It also allows collaboration by uploading and sharing learning resources (for example articles, notes, images, PowerPoint files, etc). Moreover, students can use these facilities to build upon their existing knowledge and create new knowledge through online debate and discussion (Britain and Liber, 1999; Milligan, 1999). Additionally, VLE can link directly to other systems in the institution such as, institution's library system. Students log in once to the VLE (using user name and password) can move between one systems to another without having to log in again (JISC infoNet, 2004). Therefore, the frequency of students access institution's library and other online resources can be tracked and this information can be used to measure student's engagement in their academic life.

At the same time, while providing supports and facilitates to students learning and teaching, VLE store numerous important data about the learner and the learning process. For example, VLE provide information about how often and when students have accessed a VLE, when and what students have read in the online discussion area. They also provide information about students' assignment such as the submission date and time. This information can be used in measuring students' engagement in their academic life. It is claimed that students who are actively engaged with their studies will visit the VLE (Blackboard, Moodle) more frequently and also spend longer periods of time than the students who are less engaged in their study (Beer *et al.*, 2010). Moreover, it is asserted that the students who spend more time in their study or institutions are more engaged in their study/institution and therefore, they are more likely to stay in their programme of study or institutions (Astin, 1984).

Furthermore, students often complain that they are not marked on their efforts in a group work. Information about how much work they have contributed in a group work can be collected from online discussion and student presentation area in VLE and this information can help tutors to provide fair marking (JISC infoNet, 2004). Blackboard[54] and WebCT[55] are two most popular VLEs currently being used in the UK HE (O'Leary, 2002).

### 5.2.5 Resource Information

This repository contains institution's educational settings such as, the buildings, classrooms, laboratories, libraries, studios of the campus, residential halls, facilities, equipment, supplies, and so on. Institution's educational setting information need to be made accessible across departments in the institution to support student retention (Tinto, 2000). This information can also support to attract local/international students if institutions make it accessible outside of the institutions. Also, HEI can make this information available to the accreditation bodies for efficient accreditation, as the resource information is considered in the accreditation process for quality assurance of the institutions (Eaton, 2009). Moreover, sharing this resource information across departments will help to minimize higher education institutions cost by sharing these resources across departments (Hirsch and Weber, 1999)

### 5.2.6 Research Information

#### 5.2.6.1 Research output

According to Prime Minister's Strategy Unit (2007), Department for Innovation Universities and Skills (DIUS) (2008) and Higher Education Funding Council for England (HEFCE) (2009b; 2009a), for collaborating globally in research and to strengthen the research capacity and improve the quality of the higher education institutions, the institutional repositories that contain research output needs to be shared outside of the institutions (across institutions, industries etc.) so that institutions can know each other's research works and

[54] http://www.blackboard.com/

[55] http://www.webct.com/

can take initiative for future collaborative project with other institutions. Repositories that contain research output need to be made accessible to industries or funding bodies for commercialization of research to contribute to the social economy (Hirsch and Weber, 1999; DIUS, 2008; HEFCE, 2009a; 2009b). Institutions can also attract funding bodies by visualizing their research output. Research output need to be made accessible to the accreditation bodies as research is one of the key factors to be considered in the accreditation process (Hirsch and Weber, 1999; Bridges, 2000; Eaton, 2009). Research is one of the criteria to get funding from HEFCE in research of HEIs and this funding is based on the assessment of Research Assessment Exercise (RAE) in the UK. From 2015-16, this funding will be based on the assessment of Research Framework Excellence (REF), which is a replacement of RAE in the UK. Moreover, to efficiently support critical thinking and argumentation research output need to be made accessible across institutions (Duffy *et al.*, 1998; Anderson *et al.*, 2001). It is believed that this repository also needs to be made accessible across institutions for efficient construction of personal and group knowledge.

#### 5.2.6.2 Research project

A repository that contains new areas of research and current research project, mainly those with high market demand, need to be made accessible across institutions, industries, and business to compete and collaborate globally in research (BIS, 2009; HEFCE, 2010a; 2010b). If institutions know each other's research, they would collaborate research projects of similar interests. HEFCE and QAA encourage HEI to research collaboratively. Also, the repository of research project needs to make accessible to industries, business or other funding bodies to attract funding (Hirsch and Weber, 1999; BIS, 2009; HEFCE, 2009b).

#### 5.2.6.3 Research data

Repository that contains research data needs to be made accessible only to a specific group of people. Research data can be made accessible to the members of collaborative projects so that they can update and communicate accordingly. We believe preserving research data will enhance the quality of research as well.

### 5.2.7 Academic staff, Research staff and Expertise Information

Repository that contains general (age, ethnicity, etc.), academic (qualifications, etc.) and skills information of academic staff, research staff and experts needs to be made accessible across institutions to efficiently support critical thinking and argumentation by providing relevant information (Duffy *et al.*, 1998). This information also needs to be made accessible across institutions to support cross-curricular activities by matching people and resources in the emerging area (Bridges, 2000; Tiropanis *et al.*, 2009d). Moreover, research staff and expertise information needs to be made accessible across institutions, industries, and business for collaborating globally in research and to strengthen the research capacity (Hirsch and Weber, 1999; Bridges, 2000; BIS, 2009; HEFCE, 2009b). Hence, improving the overall quality of research in the institutions as the best researchers can work together in the same research areas of their expertise. This information can also support in personal and group knowledge creation as people can easily find out people in the same area. Quality Assurance Agency (QAA) and Higher Education Funding Council for England (HEFCE) believe that to manage staff (new recruitment, tenure) efficiently higher education governance and management need this information available.

This information needs to be made accessible to the accreditation bodies for accreditation of the institutions as this information is considered in the accreditation process (Ounnas *et al.*, 2006). This information is also considered for distribution of funding in the higher education institutions (Education and Skills Committee, 2003; HEFCE, 2010a; 2010b).

### 5.2.8 Accreditation Records

Repository that contains potentially specific accreditation data needs to be integrated which is scattered across departments and needs to be made accessible to accreditation bodies for accreditation of higher education institutions (Hirsch and Weber, 1999; Hanna, 2003; Tiropanis *et al.*, 2009d). This type of repository may contain information about student support services, fiscal and administrative capacity, recruitment and admission practices, record of student complaints etc.

Table 5.1 Institutional repositories relate to higher education challenges.

| **Institutional Repositories** | **Higher Education Challenges** |
|---|---|
| Course information | Curriculum design/alignment, Fair admission and widening participation, Student employability, Quality of learning and teaching, Accreditation of HEI and programme, HE funding. |
| Teaching and learning material | Quality of learning and teaching and Student retention, progression and completion. |
| Student admission data | Group formation for learning and teaching, Student retention, progression and completion, Fair admission and widening participation. |
| Student academic record | Group formation for learning and teaching, Student retention, progression and completion, Student employability. |
| Virtual learning environments | Group formation for learning and teaching, Student retention, progression and completion, Collaborating in research, Critical thinking and argumentation, Construction of personal and group knowledge, Assessment and Feedback. |
| Resource information | Student retention, progression and completion, Fair admission and widening participation, Accreditation, Minimizing cost of HEI. |
| Research output | HE funding, Accreditation of HEI and programme, Collaborating in research, Quality of research, Critical thinking and argumentation, Contribution to economy, Increased engagement with industry, business and wider Community, Construction of personal and group knowledge. |
| Research project | HE funding, Collaboration in research, Accreditation of HEI and programme. |
| Research data | Collaborating in Research, Quality of research. |
| Academic, research staff and expertise information | Cross-curricular initiatives, Collaborating in research, Quality of research, Accreditation of HEI and programme, Critical thinking and argumentation, Construction of personal and group knowledge, Higher education leadership and management, HE funding. |
| Accreditation records | Accreditation of HEI and programme. |
| Training information | Developing new generation of staff. |
| Staff facilities | Tenure, Accreditation of HEI and programmes. |

### 5.2.9 Staff Facilities and Development Programme

#### 5.2.9.1 Training information

Repository that contains available training information for faculties needs to be made available across the departments to manage and develop new generation of staff in the institutions (Hirsch and Weber, 1999). QAA and HEFCE also state that teachers and other staff in the institutions can develop their knowledge and skills whenever they find it is necessary for their career. This way new generations of teachers will be more supportive to their students (West, 1999).

#### 5.2.9.2 Staff facilities

This repository contains information about alternative occupation, flexible age-of-retirement scheme and other facilities for staff. Therefore, this repository needs to be made available across the department in the institutions to support tenure and staff management in the institutions. For example, if a department is shut down or if the quality evaluation of teaching and research is insufficient or senior faculty those perceived as no longer productive should take into careful consideration by providing them alternative facilities (Hirsch and Weber, 1999). So that upcoming students will not fear to pursue an academic career in future. Also, this information needs to be made accessible to specific interested bodies like QAA for quality assurance of the HE institutions.

## 5.3 External Open Data Sources/Repositories

In the UK, a number of external bodies regularly publish Open Data (OD) on the Web, which could be used in addressing a number of higher education challenges. For example, student retention, progression and completion can be supported through integrating this data to other datasets and analyzing the new set of data. From the literature review on student retention, progression and completion, we found researchers traditionally use questionnaires to collect student data. The external bodies publish some similar data in their open data repositories, which could be used instead of those traditional questionnaires. Table 5.2 presents some example of traditional questionnaires, which can be replaced by using external open data sources. In the following, we describe some open data sources/data repositories that can be used in addressing higher education challenges:

POLAR2 (Participation of Local Areas) is a classification of small areas across the UK, showing the participation of young students in HE for geographical areas. This classification shows how the chances of young students entering HE vary by where they live. There are five young participation quintile groups (qYPR) of areas ordered from '1' (those wards with the lowest participation) to '5' (those wards with the highest participation). This data file can be found in ***Higher Education Funding Council for England (HEFCE[56])*** website. Students have been allocated to their neighbourhoods on the basis of their postal codes. According to HESA, those students whose postal code falls within quintile 1 are denoted as being from "low participation neighbourhoods" and those falls within quintile 2 to quintile 5 are denoted "other neighbourhoods" and all postcodes with unknown quintiles are denoted as "unknown". Therefore, it is possible to group students as low participation group and other participation group by using their postcode of their permanent address from the student admission dataset.

- ***The Office for National Statistics (ONS***[57]) publishes annual income by profession/occupation. Moreover, social-economic class (SEC) information can get from ONS published dataset. ONS classify SEC based on the occupation. These open datasets can be used to derive students' parental income as well as socio-economic status. As parent's occupation information can be attained from the student admission dataset, it is possible to link these occupations to the ONS occupation dataset to get the parental annual gross income and socio-economic class of the students.
- Every year the ***National Student Survey (NSS)*** is conducted to measure students' satisfaction in teaching and learning, assessment and feedback, academic support, organization and management, learning resources and personal development. This survey results publish on ***Unistats***[58] web site. The traditional questionnaire about student's academic and

[56] http://www.hefce.ac.uk/whatwedo/wp/ourresearch/polar/polar2/

[57]http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-250731

[58] http://unistats.direct.gov.uk/

intellectual development, academic support and satisfaction on teaching and learning can be replaced by the NSS data to develop student predictive model.

Table 5.2 Example of traditional questionnaires that can be replaced by external open data sources.

| Questions | Replacement | External data sources |
|---|---|---|
| What is your mother's annual income? | | ONS |
| What is your father's annual income? | | ONS |
| What is your socio-economic class? | | ONS |
| I am satisfied with the opportunities to meet and interact informally with faculty members. | I have been able to contact staff when I needed to. | Unistats (NSS) |
| Most faculty members I have had contact with are genuinely interested in teaching. | Staffs are enthusiastic about what they are teaching. | Unistats (NSS) |
| I am satisfied with my academic experience at this university. | Overall, I am satisfied with the quality of the course. | Unistats (NSS) |
| Few of my courses this year have been intellectually stimulating. | The course is intellectually stimulating. | Unistats (NSS) |
| My academic experience has had a positive influence on my intellectual growth and interest in ideas. | As a result of the course, I feel confident in tackling unfamiliar problems. | Unistats (NSS) |
| I am satisfied with the extent of my intellectual development since enrolling in this university. | The course has helped me present myself with confidence. | Unistats (NSS) |
| I am more likely to attend a cultural event now than I was before coming to this university. | My communication skills have improved. | Unistats (NSS) |
| Student from which participation neighbourhoods? | | HEFCE |

- ***The Higher Education Statistics Agency (HESA***[59]) publishes lots of information about UK higher education such as students' non-continuation rate by age marker, previous HE marker, low participation marker, entry qualification and subject of study. This information can also be utilised in retention study.

## 5.4 Opportunities and Challenges in Sharing Repositories

Institutional repositories (IR) are gaining popularity. This can be recognised with the adoption rates of repositories in the higher education institutions. According to Tiropanis *et al.* (2009d), over forty universities are reported to employ repositories in the UK higher education or further education to publish their research output, conference and journal articles, presentations or course material. Institutional repositories, by capturing, preserving, and disseminating a university's collective intellectual capital, serves as meaningful indicators of an institution's academic quality. IR is very handy for maintaining a collection of works as well as for preserving future use. There is a value to be gained by letting institutions have access to external repositories and by sharing their data with them. Sharing repositories have many advantages and exposing data for sharing can provide significant value in addressing higher education challenges and in supporting teaching and learning activities. Morrison (2006), Raym (2006) and Tiropanis *et al.* (2009d) stated some of the advantages in sharing institutional repositories, which are:

- Institutional repositories provide an easy way to share works.
- Interoperable repositories support the researcher's ability to search seamlessly across repository types, facilitating interdisciplinary research and discovery of new research.
- Learners and teachers will be able to efficiently search across various repositories. Learning and teaching will be better supported with utilities that enable targeted searching on authoritative teaching and learning material across institutional repositories.

[59] http://www.hesa.ac.uk/index.php?option=com_content&task=view&id=2064&Itemid=141

- Institutional repositories complement existing metrics for judging institutional productivity and prestige where sharing repository enables way to attract funding from both public and private sources.

At the same time, there are certain challenges in sharing repositories that need to be adequately discussed and addressed. According to (Lynch, 2003; Morrison, 2006; Davis and Connolly, 2007; Kim, 2007; Dietze *et al.*, 2013) there are some problems in open access and reasons for not using and sharing institutional repositories. Some of them are:

- In different data sources, data is in different formats. Therefore, interoperability becomes an issue.
- Concerns about redundancy with other modes of disseminating information.
- Confusion with copyright.
- Fear of plagiarism and having one's work scooped. Publishing someone’s work (e.g. article) before formal publication is an afraid of unscrupulous use of data and results.
- The perception of open access content being of low quality while quality is big concern of reputation for any academics.
- A lack of mandatory policies for depositing manuscripts.
- Confusion and uncertainty about intellectual property issues.
- Concern about scholarly credit and how the material in institutional repositories would be used.
- Research/teaching materials on publicly accessible web sites are not preserved in perpetuity and also they are not maintained securely.
- Publishers’ policy is another factor as they do not allow posting pre-or-post refereed articles on publicly accessible web sites and
- Additional time and effort is required to make materials publicly accessible on the Internet.

In order to potentially respond to the higher education challenges by linking and sharing institutional repositories, the above issues need to be documented properly to enhance our understanding on the pedagogical potential of institutional repositories. We need to take necessary steps to solve the above concerns relating to linking or sharing institutional repositories to get the greatest benefit from them in the higher education institution. According to (Klump *et al.*, 2004; Rae, 2007; Tiropanis *et al.*, 2009c; Dietze *et al.*, 2013), some of the approaches to address the above problems are:

- Expose the institutional repositories following linked data principles. In this way, it will facilitate the data interoperability as it allows exposing data in a standardized and accessible way. To facilitate interoperability, one of the particular strengths of the linked data approach is that linked data does not impose common and shared schemas but instead, accepts heterogeneity and offers solutions by fundamentally relying on links between disparate schemas and datasets (Dietze *et al.*, 2013).
- Data publication needs to offer authors an incentive to publish data through long-term repositories.
- Data publication requires an adequate licence model.
- Data need to be anonymized before exposed/sharing to any third party in order to protect personal information.

## 5.5 Summary

In this chapter, we discourse the data and data sources that relates to addressing the higher education challenges and group them into two broad categories: institutional repositories and external open data sources/repositories. We grouped institutional repositories into 9 key repositories containing: course information, teaching and learning materials, student records, virtual learning environments, resource information, research information, academic staff, research staff and expertise information, accreditation records and staff facilities and development programme details. Each repository addresses more than one of the higher education challenges, and to do so much of the data held in the repositories need to be shared inside and outside of the institution. In order to understand the full potential of institutional repositories in addressing the higher education challenges, we also discussed the opportunities and challenges in sharing repositories. Finally, we discussed approaches to address those challenges of sharing repositories.

This research focuses on addressing student retention, progression and completion with the available institutional internal data and external open data. Using the available data, this research develops predictive models to predict students who are at-risk to fail in their programme of study. Using this process, higher education institutions can identify poor performing students within shorten time. Therefore, HEI can arrange additional support for those poor performing/at-risk students to success without doing too late before their final exam.

This research sought to overcome the problem of traditional survey based research, which takes long time to complete and apply its result into practical context. As this research approach uses only the available data, it takes shorten time to complete. Therefore, it can be said that the process used in this research can be used in any other domain such as medical, business marketing or other initiatives where the researchers requires shorten time-frame to complete and apply the research results into practical applications.

# Chapter 6: Experimental Design and Methodology

This chapter presents the experimental platform and approach employed to examine the sufficiency of linked data technologies and linked open data sources to support student retention, progression and completion. This is undertaken through the development of student predictive models that predicts students' likelihood of being at-risk and their performance/marks in their first year of study.

## 6.1 Introduction

As discussed in Chapter 3, a diverse range of work has been undertaken in the area of student retention, progression and completion. Researchers have evaluated retention from a student standpoint as well as from an institutional standpoint. Early studies laid the theoretical foundation for scholarly inquiry into the host of factors that influence student enrolment persistence and degree completion (Tinto, 1975; Bean, 1980; Pascarella and Terenzini, 1980; Pascarella *et al.*, 1983; Astin, 1984). From a student standpoint, there are many variables that influence the likelihood of progression and degree completion, namely: high school academic achievement, parents educational qualification, socioeconomic status, gender, commitment to earning a degree, and social and academic involvement (Pascarella *et al.*, 1983; Tinto, 1993; Browitt and Walker, 2007; Miller and Herreid, 2008). In particular, it is well known that students who are less likely to persist through their studies are typically socially disadvantaged, academically less prepared, and who experience a lack of resources and support from significant others (Braxton, 2000; Seidman, 2005). We also know that those who feel isolated or lack a sense-of-belonging during their early years of study are more likely to leave their programmes or institutions (Hurtado and Carter, 1997; Hausmann *et al.*, 2007). However, there are few papers that engage with easily accessible data and standard empirical techniques in order to identify students who at-risk for prematurely terminating their studies (Murtaugh *et al.*, 1999; Miller and Tyree, 2009; Singell and Waddell, 2010). As discussed in Chapter 3, it has been recognised that most of the studies on student retention, progression and completion are often survey-based, where researchers used questionnaire to collect students' data to analysis. Survey-based studies have some drawbacks such as low participation rate, high cost.

The primary goal of this research is to determine if combining institutional internal datasets/repositories and external open data sources/repositories can provide accurate and improved predictive models to monitor student retention. This enquiry would allow institutions to develop student predictive models to monitor student progression without having to rely on traditional questionnaires. Moreover, this research will examine whether linked data technologies is sufficient to build student predictive models by combining data from institutional internal and external open data sources/repositories.

In the following sections, two experiments have been designed to test the hypotheses of this research as presented in Chapter 1. The first experiment is designed to predict students who are at-risk in their first year of study. The second experiment seeks to predict students' academic performance/marks in their first year of study based on easily accessible data from institutional internal data sources/repositories and external open data sources. An online questionnaire was conducted to collect data for both experiments. This questionnaire was used to collect data available through the institutional databases and in the traditional questionnaires. This information was combined with other external open data in an effort to predict students who are likelihood to be at-risk of leaving their programme/institution and students' academic performance/marks in their first year of study. Linked data technologies were applied in combing data from different data sources.

Section 6.2 explains the linked data experimental platform with developing an ontology, which has been developed in conducting the experiments to combine datasets from different data sources and to make the final set of data to further analyses. Section 6.3 discusses the planned experiments, whereas section 6.4 and 6.5 describe the participants and the survey used to collect student specific data for the experiments. Finally, section 6.6 summarises the key finding from this chapter.

## 6.2 Experimental Environment/Platform

As data exists within different data sources (e.g., institutional internal and external), the biggest challenges and opportunities lie in connecting these disparate datasets together in order to create a single set of integrated data for analysis. Combining data into a common location is inhibited by different technology standards, lack of unique identifiers, and organizational challenges to the ownership and use of the data (Arnold, 2010). Linked data is well suited for data integration while data is in different formats in different data sources. Therefore, we develop a link data infrastructure to examine the sufficiency of linked data technologies to develop predictive models to support student progression. This involves integrating related data from disparate data sources (institutional internal or external) and analyzing the new set of linked data. Figure 6.1 depicts our vision in developing student predictive models through integrating data from disparate data sources.

We outline the following requirements are important for the experimental platform to build the student predictive models by combining data from institutional internal and external repositories.

i. Understand the concepts and relationships that exist in the student retention, progression and completion domain.

ii. The ability to convert raw data to Resource Description Framework (RDF): Currently most of the interested datasets are not in linked data format. They are in different formats (e.g., .csv, .xls). Linked data principle entails to share and follow a common data format and RDF is a standard for representing linked data. RDF offers many advantages, such as provision of an extensible schema, self-describing data, de-referenceable URIs, and, as RDF links are typed, combinings of different datasets are easy and safe. Therefore, the platform desires to have the ability to convert data to RDF.

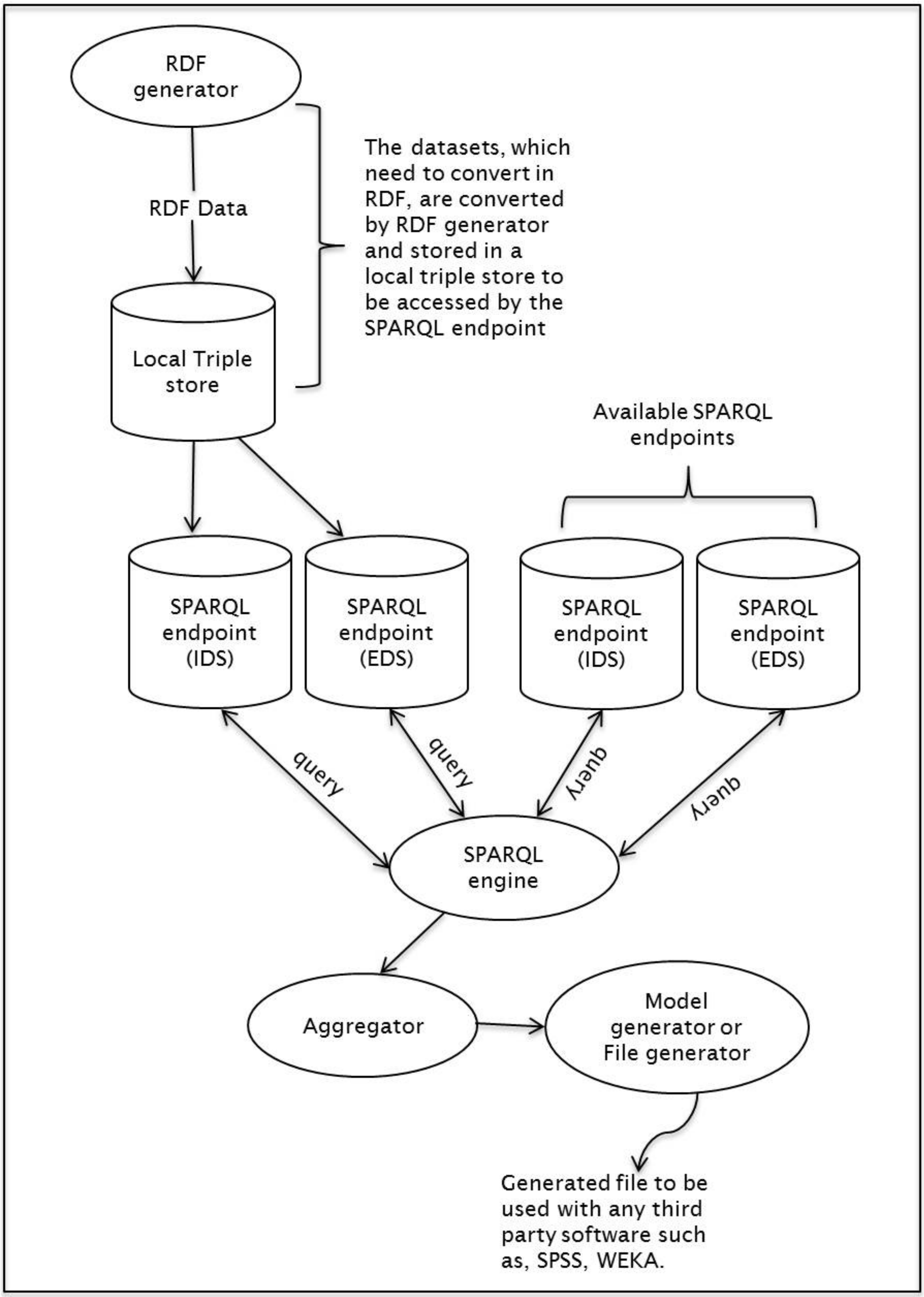

RDF generator
RDF Data
Local Triple store
The datasets, which need to convert in RDF, are converted by RDF generator and stored in a local triple store to be accessed by the SPARQL endpoint
Available SPARQL endpoints
SPARQL endpoint (IDS)
SPARQL endpoint (EDS)
SPARQL endpoint (IDS)
SPARQL endpoint (EDS)
query
query
query
query
SPARQL engine
Aggregator
Model generator or File generator
Generated file to be used with any third party software such as, SPSS, WEKA.

Figure 6.1 Experimental platform.

iii. The ability to perform SPARQL query over different RDF sources. SPARQL is used to express queries across diverse RDF sources.

iv. The ability to join multiple SPARQL query results into a single dataset for data analysis as predictive models will be developed on the final integrated dataset.

v. The ability to develop the predictive model or the ability to save the final dataset into a file to use by other third party software for further analysis.

From the above requirements, the first step is to understand the datasets and the relationship between the data and datasets related to student retention, progression and completion. Specifically in this step, an ontology is developed to structure data and states the links to join the datasets. Section 6.2.1 outlines the ontology developed to structure the data and define the links exists in the datasets. The subsequent sections 6.2.2, 6.2.3, 6.2.4, 6.2.5 inclusive, fulfil the remaining requirements by contouring four functional components necessary for the linked-data based experimental platform.

### 6.2.1 Modeling the Datasets

The fourth principle of linked data encourages data providers to join their datasets to others by explicitly stating the relationships between the data they publish and the data already published by third-parties. In this way, the ontologies play a central role to our research since they are used to structure data and their relationship and thus, their capacity to be combined. An ontology defines a vocabulary to model a domain, as well as, a set of explicit assumptions regarding the intended meaning of these terms (Guarino, 1998). Thus, the data obtains a formal “meaning”. Among several definitions of ontology proposed (Gómez-Pérez *et al.*, 2004), the most common defines an ontology as “a formal specification of a share conceptualization” (Borst, 1997). Allemang and Hendler (2008) propose three ontology languages to model and implement ontologies, all of which are based on RDF:

- Simple Knowledge Organization System (SKOS) is a data model to express conceptual hierarchies. It is a lightweight ontology that provides enough expressiveness to describe vocabularies or taxonomies. For this reason, SKOS is used by those data providers who do not need to implement complex relationships and prefer their knowledge bases to keep simple.

- RDF Schema (RDFS) is a widely used vocabulary to implement lightweight ontologies. It can be seen as a natural extension of RDF that includes a vocabulary to express classes, individuals, properties and taxonomies. Thus, ontology concepts can be express although its inference capability is still limited.
- Ontology Web Language (OWL) is another language to implement ontologies that extends the expressiveness of RDFS. It is a much more complex ontology language that enables to express axioms and restrictions. This way, relationships between classes, properties and individuals can be formalized; besides, OWL-based ontologies are typically very complex and hard to be reused.

There are two different ways of formalizing the relationships in the datasets. The first possibility is to reuse vocabularies from other datasets to describe data as seen in OWL (Bechhofer *et al.*, 2004). The second possibility is to include the relationships between the data published and the data contained in other datasets. Therefore, datasets in the Web of Data are linked and a data consumer application can automatically discover, access and integrate data from different data sources (Heath and Bizer, 2011). In the proposed research, develop a simple Student Retention, Progression and Completion (SRPC) ontology by defining key concepts and explaining the relationships between them. This ontology is not a complete ontology; rather, it mainly describes the factors as stated in the literature that those influence students' likelihood to remain or dropout from their study. At present, it consists of students' background information (e.g., ethnicity, neighbourhood, entry route, accommodation type, employment status, disability status) and academic information (e.g., entry qualification, semester marks). In an ontology, classes are the central element and define the categories to structure the information. Ontologies usually organize classes into taxonomies where inheritance mechanisms can be applied. Each class has a set of properties, which may have some restrictions. Figure 6.2 presents a simplified version of SRPC ontology, whereas, *Appendix D* presents all the classes and properties of the SRPC ontology. We foresee that in the future the ontology's vocabulary will evolve to include more classes and properties. Some of the vocabulary used here may also be changed in order to ensure a better understanding of the concepts behind the terminology. For example, not all educational systems agree that on using the term "module" is the appropriate vocabulary to describe a course unit.

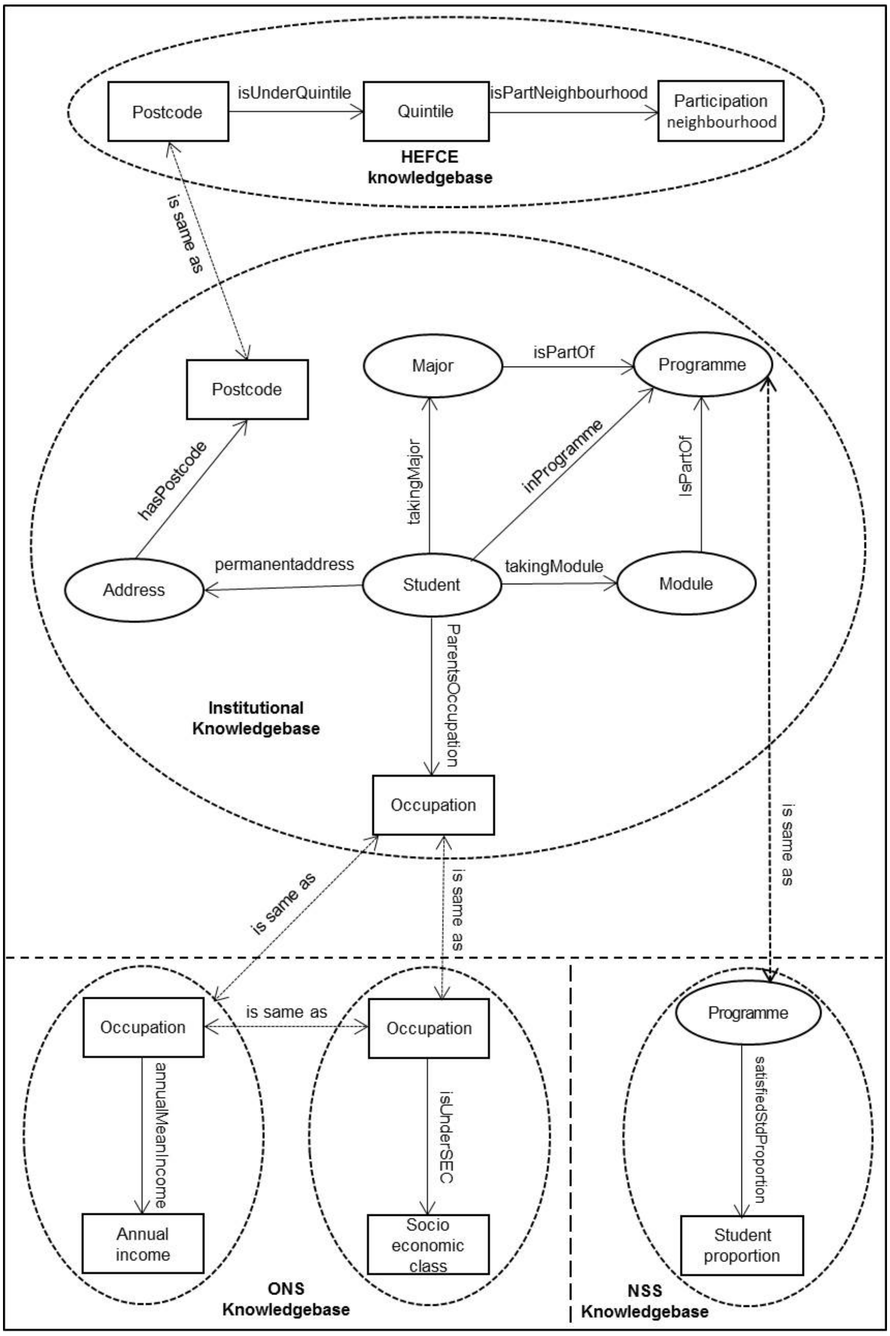


Figure 6.2 The simplified version of SRPC ontology.

Also while developing SRPC ontology, some of the vocabularies taken from Friend-of-a-Friend (FOAF[60]) vocabulary, thereby following the advice given in (Bizer *et al.*, 2007) to re-use terms from well-known vocabularies. For example, in this research foaf: age, foaf: gender and foaf: fundedBy were used from FOAF vocabulary to refer age of a student, gender of a student and source of funding of the study expenses of a student respectively.

### 6.2.2 RDF Generator

As most of the datasets of interest are not yet in linked data format, we developed a number of scripts. These scripts are able to automatically convert the datasets (.csv) into RDF triples. Besides, we used the existing tools to convert data into RDF, as needed, such as Grinder[61], google-refine[62].

### 6.2.3 SPARQL Engine

This component provisions to connect to different SPARQL endpoints. It only supports sending SPARQL queries via HTTP requests (i.e. sending queries to SPARQL endpoints) and accepts query results via HTTP as well.

### 6.2.4 Aggregator

It supports to join multiple SPARQL query results into a single dataset based on a common identifier. For example, there are two query results, result1 and result2. These query results have a common identifier: students' identification number, or otherwise known as "ID". Based on this common ID the "aggregator" joins these two datasets into a single dataset, result3. Hence, result3 = result1 U sresult2.

### 6.2.5 Model Generator or File Generator

After combining multiple RDF sets into a single RDF set, the next step is to develop student predictive model based on this single dataset aggregated from different data sources. This has the ability to save the final dataset in a file to be used in any custom written or any available software, such as R statistics, SPSS,

[60] http://xmlns.com/foaf/spec/
[61] https://github.com/cgutteridge/Grinder
[62] http://code.google.com/p/google-refine/

Rapid Miner or WEKA to develop the predictive models or any further analysis of the data.

We used eclipse[63] 3.6 and openrdf-sesame[64] version 2.6.0 in our implementation. Sesame is a generic architecture for storing and querying RDF and RDF schema, which was proposed by Broekstra *et al.* (2002). We deployed the experimental platform on an iMac with 4GB of RAM that runs a Mac operating system version 10.6.8. To do the statistical analysis on the refined (transformed) dataset IBM SPSS Statistics 20 and WEKA were used in our experiment.

## 6.3 Experimental Methodology

This section describes the experimental design and data analysis methods of the two experiments to answer the research questions mentioned in Chapter 1. The aim of the experiments is to determine whether linked data and external open data sources or external repositories can be used to develop student predictive models to support student retention progression and completion. Moreover, the experiments investigate whether external open data sources could be used instead of traditional questionnaires in developing student predictive models. Table 6.1 shows the relationship among the experimental studies, research questions and hypotheses. The following two sections, 6.3.1 and 6.3.2, describe the experimental methodology for experiment 1 and experiment 2 respectively. Figure 6.3 shows this research's methodological steps.

### 6.3.1 Experiment 1: Exploring Student Progression Predictive Models based on Institutional Internal Datasets and External Open Data Sources

#### 6.3.1.1 Experimental design

To develop the student predictive models, experiment 1 examines the sufficiency of linked data and external open data sources. Typically research in this area is carried out using questionnaires, which have few drawbacks including costs and potentially low participation rates.

[63] http://www.eclipse.org/

[64] http://www.openrdf.org/

Table 6.1 Relationships among the experimental studies, research questions and hypotheses

| Hypotheses | Research questions | Experimental studies |
|---|---|---|
| *Hypothesis 1*: It is possible to provide accurate/improved student prediction models by combining institutional internal databases and external open data sources.<br>*Hypothesis 3*: Institutional internal/external data sources can be used to compensate the lack of questionnaire data in building student prediction model. | Part of research question 2: Can we show how student retention, progression and completion can be efficiently addressed by aggregating information using linked data technologies from internal or even external data sources? | Experiment 1 and Experiment 2 |
| *Hypothesis 2*: Linked data can provide sufficient support for building student prediction model when combining institutional internal/external data sources. | Part of research question 2: Are linked data technologies well suited to address this challenge? What are the advantages of using linked data technologies in this respect? | Experiment 1 |
|  | Research question 3: Can we provide an infrastructure to efficiently monitor any potential data patterns that indicate stay/drop in student retention, progression and completion? What would be the challenges to provide such an infrastructure? | Experiment 1 |
| *Hypothesis 4*: It is possible to predict students' mark using institutional internal and external data sources. |  | Experiment 2 |

The experiment has been designed and conducted to validate the new predictive model in comparison to the survey-based model by using data available from institutional internal databases and external open data sources.

An analysis of experiment evaluates whether traditional questionnaires about students' academic and intellectual development, academic support and satisfaction on teaching and learning can be replaced by an externally provided National Student Survey (NSS) questionnaires.

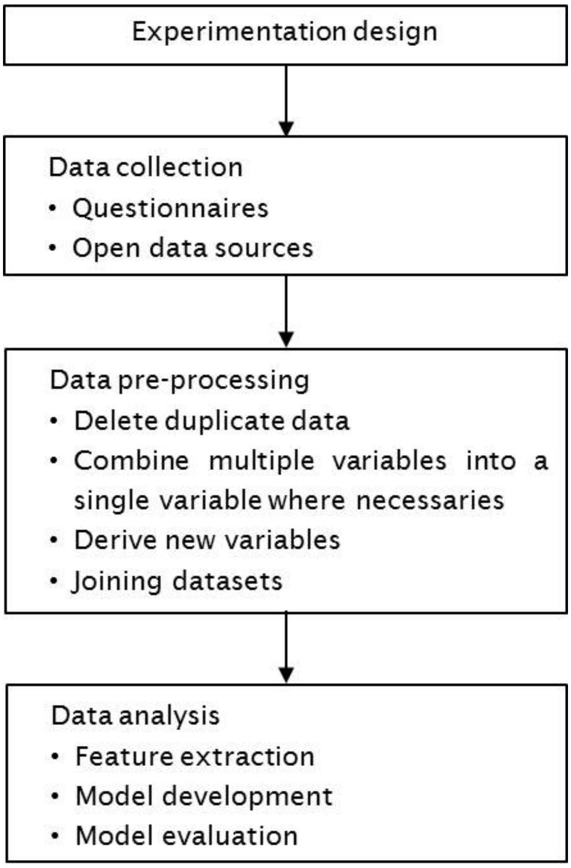


Figure 6.3 Experimental methodologies.

This experiment designs and compares three models to predict students who are likely to be at-risk in their programme of study based on previous work. The first model includes all the independent variables considered by Pascarella *et. al* (1983) in their study of predicting dropout students. Their survey-based study is derives from Tinto's student retention theory (1975). The second model is developed including only the variables from the first model (survey-based model), which are commonly available in the institutional internal databases. Finally, the third model includes all the variables from the second model and includes new variables from external open data source as the replacement of the traditional questionnaire items/variables from the first model.

#### 6.3.1.2 Data collection

In order to collect data for the experiment, a survey was conducted in June 2012. The targeted participants were first-degree/undergraduate students who have been enrolled in the academic year 2010/2011 in any programmes of study at the university of Southampton. As some sensitive and personal data were collected through the survey such as, ethnicity of the students, the university's Ethics Committee approval was required before the start of the survey. The university's Ethics Committee approval was obtained with Ethics id 1978 on18th May 2012 (*see Appendix B for reviewed documents*).

As during that time we did not have permission to access institutional internal databases, we conducted the survey to collect data for our experiment. The survey itself is a questionnaire comprising of 49 questions and divided into six parts (see *Appendix A* for details), and expected completion time was about 20-30 minutes. The aim of this survey was to collect institutional internal databases item as well as to collect traditional questionnaires data. The questionnaires used in this survey were largely based upon Tinto's retention model as from the literature we found Tinto's retention model is the key in the area of student retention, progression and completion. Also, Pascarella and Terenzini's constructed Institutional Integration Scale (IIS) was used to measure all the constructs of Tinto's retention model such as peer group interaction, student-faculty interaction, faculty concern for student developing and teaching, academic and intellectual development, goal commitment and institutional commitment. Moreover, in this research, IIS serves as the baseline in replacing questionnaire data with the available external open data.

In order to facilitate and speed up the process, the questionnaire was provided online. The questionnaire was available to participants between 01/06/2012 and 30/08/2012 at https://www.isurvey.soton.ac.uk/5072. The students who participated in the survey were also invited to take part in a prize draw of 20-pound amazon voucher as an incentive for their valuable participation. To participate in the prize draw, students had to provide their university email address. However, this email address was kept separately from the questionnaire results, so that the survey was completely anonymous. Apart from the survey, we used National Student Survey (NSS) open dataset in this experiment.

#### 6.3.1.3 Data pre-processing

The data collected through survey was not in the format ready for an easy and direct analysis and modelling. Therefore, data pre-processing was required to prepare database for modelling. In this step, data were cleaned for any duplication of records. For example, when we found more than one record under the same student email id then we kept one record and deleted the rest. Also, multiple variables with small proportion of data are combined into a single variable to increase the model parsimony.

In *Appendix E*, provides all the variables used in this experiment with their associated domain value for reference. We highlight here the domain values for some of the variables defined for this experiment:

- **Age:** Age is a numeric continuous variable, which was converted into a categorical variable with only two age groups: young and mature. Young students are those who are <21 years and the rest are in mature age group. We deleted this variable to include in our experiment, as few students were found in mature age group.

- **Gender:** The sex of the students, grouped into two groups: male and female.

- **Ethnicity:** The culture/ethnicity of the students, grouped into two groups: white and non-white.

- **A Level tariff points:** Students grade in A level of study. Grades are assigned to all students using the following mapping: A*=140, A=120, B= 100, C=80, D=60 Example, if a student's A level grades are AAA then his A level tariff points counted as AAA=120+120+120=360.

- **Accommodation Type:** Students' living place during their first year of study. Data were collected in 6 categories of accommodation through the online questionnaire, such as university hall, private halls, parents house, own residence, rented accommodation and other. During this stage, data are classified into two groups: university halls and others. Data with private halls, parents house, own residence, rented accommodation and other accommodation types are combined into a single category due to the small proportion of the data (less than 10%). Combining them into one accommodation group helps with model parsimony.

- **First year's first semester marks:** Students marks in their first semester of first year of study. Marks are assigned to all students using the following mapping: 71%-100%=1, 61%-70%=2, 51%-60%=3, 41%-50%=4.

- **Source of tuition fee:** Students' source of tuition fee. Data were collected through the online questionnaire into 5 categories: yourself, family, grant/scholarship, student loan and others. Data with "family" and "yourself" categories are combined into one group as they have smaller proportion of the data and there was no data in the "other" group. Finally, data are classified into three groups: grant/scholarship, student loan and family/yourself.

- **Study field:** Study field of the students are classified into two groups: applied and non-applied. Students who study Science, Technology, Engineering and Medicine (STEM) are into applied study group and the rest are into non-applied study group.

- **Parents' have Higher Education (HE) qualification:** This is the educational qualification of the student's parents. This variable categorized as yes/no. If any of the parents or both parents completed at least Bachelor degree then considered as "yes" otherwise considered as "no".

- **Students' working/employment status:** Students' working status in their first year of study is grouped into two groups: employed and unemployed.

- **Study outcome:** Students' subsequent academic outcome is categorised into two possible categories: at-risk and not at-risk. The students who are at-risk in their programme of study are determined based on 2 criteria: a) the students who failed to progress according to their academic year or semester that means if students enrolled in October 2010, then they were expected to be in their second year second semester at the time the questionnaire was conducted but if they are behind their expected year and semester of their study then they were identified as at-risk students and b) the students who got less than 50% marks in their first year or in their first year's first semester exam are also identified as at-risk students. On the other hand, the students who successfully progressed according to their academic year or semester are defined as not at-risk students. Also, the students who achieved 50% or more than 50% marks are denoted as not at-risk students.

- **NSS questionnaire:** We considered 16 questionnaire items (*see Appendix E or in Table 7.2 in Chapter 7*) from NSS dataset and the reason of including these 16 questionnaires is explained later in data and data sources section in Chapter 7. NSS measures students' satisfaction on their programme of study in a 5-points scale (i.e., Definitely Disagree, Moderately Disagree, Neither Agree nor Disagree, Moderately Agree, Definitely Agree). Unistats publishes the percentages of respondents in each scale for an individual course. We considered the actual value (% Agree) of the respondents for those 16 questions of 2010-2011 academic year's published result of each individual course for the university of Southampton. This field shows the proportion of students who "agree" or "strongly agree" with the NSS questions.

#### 6.3.1.4 Data analysis

This experiment uses categorical principal component analysis (CATPCA) to determine the importance of the prediction variables for modeling the study outcome. Logistic regression (LR) was used in developing the predictive models to predict the students at-risk in their study and repeated hold out method was used to validate the models. Sections 6.3.1.4.1, 6.3.1.4.2 and 6.3.1.4.3 describe CATPCA, LR and repeated hold-out method respectively. Finally, the developed student predictive models were compared based on sensitivity, specificity, type I error rate and type II error rate of each model to determine the best performing model among them.

Sensitivity is the ability of a model to detect the positive instances (at-risk students) and is defined as the ratio: sensitivity = TP/(TP+FN), where, TP (True Positive) is the number of positive instances (at-risk students) correctly classified by the model and FN (False Negative) is the number of positive instances (at-risk students) misclassified as negative instances (not at-risk students). On the other hand, specificity is the ability of a model to detect the negative instances (not at-risk students) and is defined by the ratio: specificity = TN/(TN+FP), where TN (True Negative) is the number of negative instances (not at-risk) correctly classified by the model and FP (False Positive) is the number of negative instances misclassified as positive (at-risk).

Type I error (false positive error) occurs when a student who is actually a not at-risk student is wrongly classified as the student at at-risk and is defined as, 1-sensitivity, whilst, Type II error (false negative error) occurs when a student who is actually an at-risk student is wrongly classified as a not at-risk student and denoted as, 1-specificity.

#### 6.3.1.4.1 Categorical Principal Component Analysis (CATPCA)

The goal of Principal Components Analysis (PCA) is to reduce an original set of variables into a smaller set of uncorrelated components that represent most of the information found in the original variables. The technique is most useful when a large number of variables prohibit the effective interpretation of the relationships between objects. By reducing the dimensionality, the technique interprets a few components rather than a large number of variables. Principal component is based on correlations between input variables in the R-matrix, whereby variables with similarly large correlation coefficients make up the components. Each component is then assigned a score obtained using the scores of the underlying variables with regard to the case. Using these scores, further analyses, for example correlation tests, can be carried out with factors other than the original variables.

Categorical Principal Component Analysis is the nonlinear equivalent of PCA. It aims at the same goals of traditional PCA, but it is suited for variables of mixed measurement level that may not be linearly related to each other (Linting *et al.*, 2007).

The nonlinear PCA model is beneficial as it applies to nonlinearly transformed data. The variables are transformed by assigning optimal scale values to the categories, which results in numeric valued transformed variables. The nonlinear PCA simultaneously accounts for the nature of items, the different role of items in determining the measure, and the possible multidimensionality of the concept.

PCA decomposes components using eigenvalues of the underlying variable correlation matrix. The eigenvalues denote the relative importance of the component and are used to calculate eigenvectors indicating the loadings of individual variables on that component. Following Kaiser's criterion, components with eigenvalues greater than 1 are normally retained. If the analysis has fewer than 30 input variables, this criterion is considered too strict and it is recommended to retain factors with eigenvalues greater than 0.7 instead.

Typically, the result of PCA is a table containing the component loadings of all input variables. Since these loadings denote the importance of the variable with regard to the component, they can be used to identify which input variable is part of which component. Sometimes, the interpretation of these loadings can be difficult, especially because some variables may load well on multiple factors. This can be mitigated using factor rotation, that is, let a component be a vector or axis along which variables are plotted, then the loading of variables on that component can be maximised by rotating that vector. There are different rotation algorithms that can be employed, for example, the varimax rotation aims to produce a few high valued loadings and many low-valued loadings so that the number of variables per factor/component is minimal with each variable having a maximum loading with regards to that factor/component, while quartimax rotation attempts to do the opposite.

#### 6.3.1.4.2 Logistic Regression (LR)

Most retention studies (Pascarella *et al.*, 1983; Robst *et al.*, 1998; Herzog, 2005; Miller and Herreid, 2008; Singell and Waddell, 2010) adopt a logistic regression approach. Logistic regression is an established method in retention studies for it handles both categorical and continuous predictor variables, which do not have to exhibit linearity and homogeneity of variance vis-a`-vis the outcome variable (Hosmer and Lemeshow, 2000; Peng *et al.*, 2002).

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. Logistic regression models ensure that the estimated ranges of probabilities are between 0 and 1. Logistic regression model applies a transformation to the probabilities. The probabilities are transformed because the relationship between the probabilities and the predictor variable is nonlinear. The logit transformation ensures that the model generates estimated probabilities between 0 and 1. The logit is the natural logarithm (ln) of odds of *Y*, and odds are ratios of probabilities ($\pi$) of *Y* happening (i.e., a student will stay) to probabilities $1-\pi$ of *Y* not happening (i.e., a student will not stay/drop out).

Hence, The simple logistic model has the form,

$$logit(Y) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X \cdots\cdots\cdots\cdots (1)$$

Taking the antilog of *Equation 1* on both sides, an equation derives to predict the probability of the occurrence of the outcome of interest as follows:

$$\pi = Probability\ (Y = outcome\ of\ interest\ |\mathrm{X} = \mathrm{x}, \text{a specific value of X})$$
$$= \left(e^{\alpha+\beta x}\right)/(1 + e^{\alpha+\beta x}) \cdots\cdots\cdots\cdots (2)$$

where,

$\pi$ is the probability of the outcome of interest or “event,”

ln is the natural logarithm.

α is the *Y* intercept.

β is the regression coefficient.

*e* = 2.71828 is the base of the system of natural logarithms.

Extending the logic of the simple logistic regression to multiple predictors (say *X*1 = reading score and *X*2 = gender), one can construct a complex logistic regression for *Y* as follows:

$$logit(Y) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta 1X1 + \beta 2X2 + \beta 3X3 + \cdots\cdots\cdots + \ \beta nXn \cdots\cdots\cdots (3)$$

Therefore,

$$\pi = e^{\alpha+\beta 1x1+\beta 2x2+\beta 3x3+\cdots\cdots\cdots+\beta nxn} / 1 + e^{\alpha+\beta 1x1+\beta 2x2+\beta 3x3+\cdots\cdots\cdots+\beta nxn} \cdots (4)$$

The value of the coefficient β determines the direction of the relationship between *X* and the logit of *Y*. When β is greater than zero, larger (or smaller) *X* values are associated with larger (or smaller) logits of *Y*. Conversely, if β is less than zero, larger (or smaller) *X* values are associated with smaller (or larger) logits of *Y*.

The Binary logistic regression is a form of regression used when a dependent variable takes only two values (e.g. study outcome with two values such as at-risk or not at-risk).

6.3.1.4.3 Repeated Hold Out Method

To evaluate the models, repeated hold out method was used. This method randomly splits the dataset into training dataset and test dataset. Then for each split, the model is developed using the training dataset and predictive accuracy is assessed using the test dataset. The accuracy rates on the different splits are then averaged to yield an overall accuracy rate.

### 6.3.2 Experiment 2: Students' Performance Prediction using Institutional Internal Datasets and External Open Data Sources

#### 6.3.2.1 Experimental design

Experiment 2 scrutinizes the opportunities of using external open data sources in predicting students' academic performance in their first year of study. It proposes a new method to predict students' mark based on a mix of institutional internal data and external open data. Initially, we developed two predictive models based on previous work. The first model includes all the independent variables (as many as available) considered by Yadev and Pal (2012). The second model combines the variables from the first model, which are commonly available in the institutional internal databases, with variables from external open data sources. Then, we compare the results to find out the best performing model among them. After that, students' first semester marks included in both predictive models to see the impact of first semester marks in predicting students' first year final marks. Also, it compares the results of the two new predictive models to find out the best performing model among them.

#### 6.3.2.2 Data collection

Data were collected at the same time as experiment 1's survey data collection and used the same participation. Ethics approval was obtained under the same review as the first experiment. Apart from the survey data, we used National Student Survey (NSS feedback data), Office for National Statistics (ONS) and Higher Education Funding Council for England (HEFCE) published open datasets in this experiment.

#### 6.3.2.3 Data pre-processing

In this phase, the data is put into a form suitable for analysis and modeling. At this stage, some selected variables are combined, transformed or used to create new variables as necessary.

All the variables, which were used in this experiment, are provided in Table 8.1 in Chapter 8 with their domain values for reference. The domain values for some of the variables were defined for the experiment as follows:

- **Residence:** The residence/domicile of the students, grouped into three groups: UK, Other-EU and non-EU.

- **A level points:** Students grade in A level of study. Grades are assigned to all students using the following mapping: A*=140, A=120, B= 100, C=80, D=60. For an example, if a student's A level grades are AAA then his A level tariff points counted as AAA=120+120+120=360.

- **Accommodation Type:** Students' place of residence during their first year of study. Data were collected in six categories of accommodation through the online questionnaire: university hall, private halls, parents house, own residence, rented accommodation and other. During this stage, data are classified into two groups: university halls and others. Data with private halls, parents house, own residence, rented accommodation and other accommodation types are combined into a single category due to the small proportion of the data (less than 10%). Combining them into one accommodation group helps with model parsimony.

- **Admission type:** The admission type, which may be direct admission through university procedure or clearing. Clearing[65] is an application option for students who have not received the university offer they want, who haven't got the grades they needed, or who applied to university too late.

- **First year's first semester marks:** Students marks in their first semester of first year of study. Marks are assigned to all students using the following mapping: 71%-100%, 61%-70%, 51%-60% and 41%-50%.

- **First year's final marks:** Students final marks in their first year of study. Students' first year's final marks are declared as response variable in this experiment. Marks are assigned to all students using the following mapping: 71%-100%, 61%-70%, 51%-60% and 41%-50%.

- **Father's occupation:** This is students' fathers' occupational category. Occupations are grouped into three categories: service, business and NA.

- **Mother's occupation:** This is students' mothers' occupational category. Occupations are grouped into three categories: service, housewife and NA.

- **Participation neighbourhood:** Students are categorized into their participation neighbourhood group based on the postcode of their parental/permanent address. This variable is derived linking institutional internal and external open data. Higher Education Funding Council for England (HEFCE) published participation neighbourhood dataset based on the postcode in their website. Data are categorised into three categories: lower participation neighbourhood (postcodes falls in quintile 1), other neighbourhood (postcodes falls in quintile 2 to quintile 5) and unknown (all postcodes with unknown quintiles). Figure 6.4 shows an example of deriving students' participation

[65] http://www.ucas.com/ucas-terms-explained

neighbourhood group by linking institutional internal dataset (admission dataset) and HEFCE published participation neighbourhood open dataset.

- **Parents' annual income:** Students' parents' annual gross income is derived from office for national statistics (ONS) published dataset based on their occupation. At the first instance we linked students' fathers' and mothers' occupation provided by the students through the survey to the ONS published occupational categories. Based on the occupational mapping, both parents' annual gross income was derived from ONS published dataset. Figure 6.5 shows the derivation of parents' annual gross mean income by linking institutional admission dataset and ONS published annual gross mean income for male and female.

- **Socio economic class:** This is students' socio-economic group based on their parental occupation. This variable is derived linking institutional internal dataset and external open dataset published by office for national statistics (ONS). ONS publishes annual mean gross income by occupation separately for male and female. At the first instance we linked parental occupation provided by students through the survey to the ONS published occupational categories. Based on the occupational mapping, parental annual gross income was derived from ONS published dataset. After that students' socio-economic class was derived based on the occupation of the parents who earned higher income. Socioeconomics are derived into seven classes. The full derivation table is given in *Appendix F*.

Finally, in this stage, students are classified into three socio-economic groups due to the small proportion of data in some classes. Students are grouped following ONS three class version[66] of the socio economic class: Higher managerial,

[66] http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/soc2010/soc2010-volume-3-ns-sec--rebased-on-soc2010--user-manual/index.html

administrative and professional occupations (MP-occupations), Intermediate occupations (I-occupations), and Routine and Manual occupations (RM-occupations). Table 6.2 shows the relationship between seven classes and three classes versions of the socio-economic classes (SEC), which follows to group SEC for the students.

- **NSS questionnaire:** The same 16 NSS questionnaires data used in experiment 1 such as, student faculty interaction, faculty concern for student development, students' development and about their course are also included in this experiment.

Table 6.2 Relationship between seven classes and three classes version of the socio economic class (SEC).

| Seven classes | Three classes |
|---|---|
| 1. Higher managerial, administrative and professional occupations<br>1.1 Large employers and higher managerial and administrative occupations<br>1.2 Higher professional occupations | 1. Higher managerial, administrative and professional occupations |
| 2. Lower managerial, administrative and professional occupations | |
| 3. Intermediate occupations | 2. Intermediate occupations |
| 4. Small employers and own account workers | |
| 5. Lower supervisory and technical occupations | 3. Routine and manual occupations |
| 6. Semi-routine occupations | |
| 7. Routine occupations | |

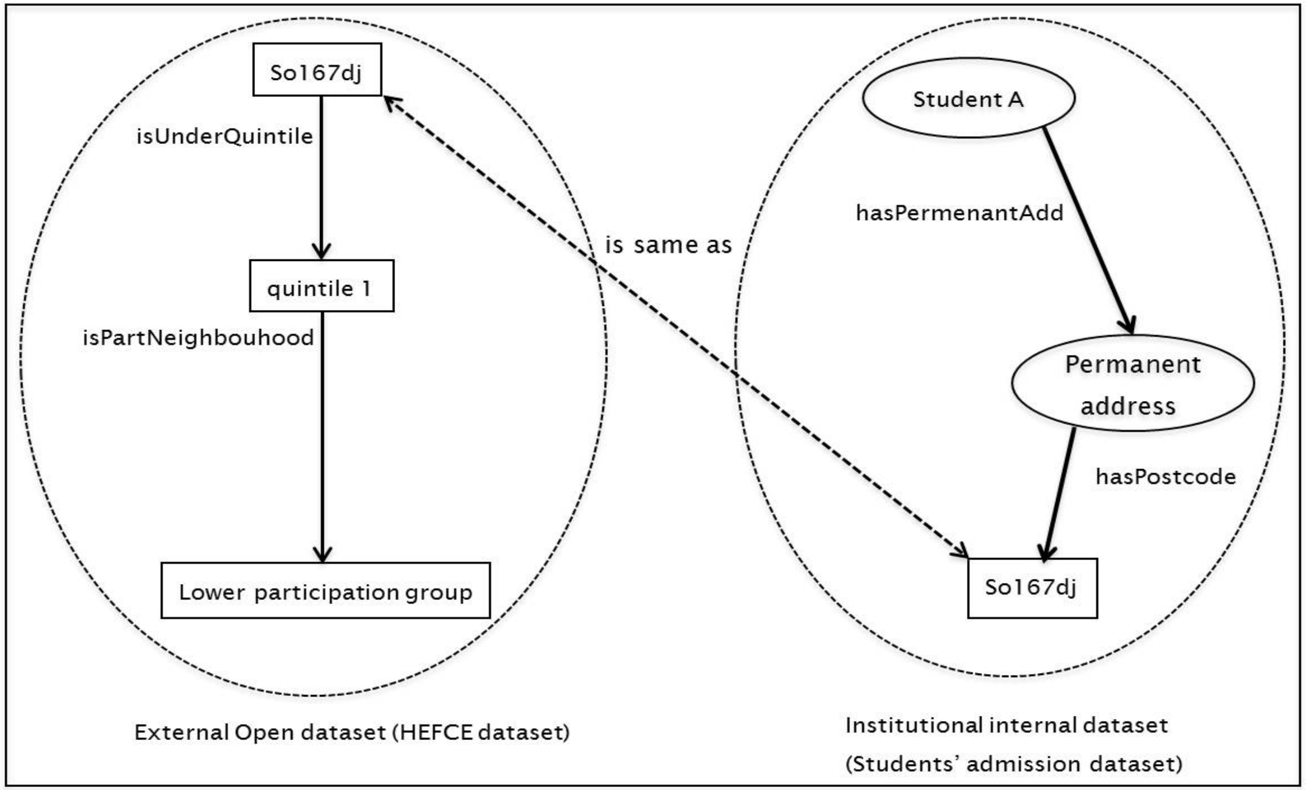

So167dj
isUnderQuintile
quintile 1
isPartNeighbouhood
Lower participation group
is same as
Student A
hasPermenantAdd
Permanent address
hasPostcode
So167dj
External Open dataset (HEFCE dataset)
Institutional internal dataset
(Students' admission dataset)

Figure 6.4 An example of deriving students' participation neighbourhood group by linking institutional internal and external open dataset (HEFCE's open dataset).

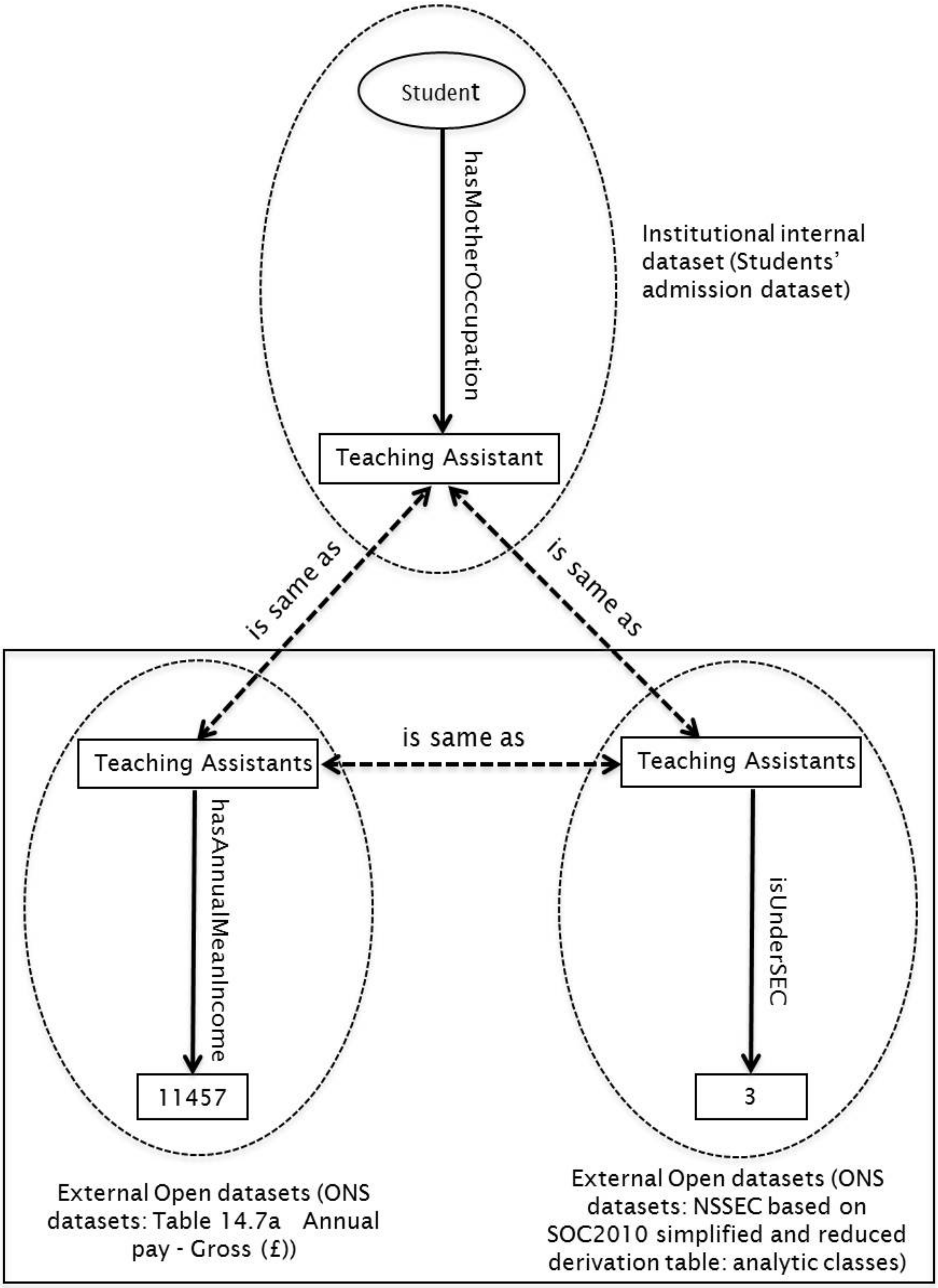

Student
hasMotherOccupation
Institutional internal dataset (Students' admission dataset)
Teaching Assistant
is same as
is same as
is same as
Teaching Assistants
Teaching Assistants
hasAnnualMeanIncome
isUnderSEC
11457
3
External Open datasets (ONS datasets: Table 14.7a Annual pay - Gross (£))
External Open datasets (ONS datasets: NSSEC based on SOC2010 simplified and reduced derivation table: analytic classes)

Figure 6.5 An example of deriving parents' annual gross mean income and socio economic class by linking institutional internal dataset and external open dataset (ONS dataset).

### 6.3.2.4 Data analysis

This experiment uses attribute/feature selection method to select the significant attributes to include in the models. Decision tree (C4.5) is used to develop the classification models and 10-fold cross validation method is to evaluate the models. Sections 6.3.2.4.1, 6.3.2.4.2 and 6.3.2.4.3 describe feature selection, decision tree and 10-fold cross validation methods respectively. Finally, to determine the best predictive model, prediction accuracy, precision, recall and F-measure of the models are measured to compare the models.

The Precision means the proportion of the instances (students) which are truly in class x among all those which were classified as class x and is denoted as the ratio: Precision = TP/(TP+FP), where, TP (True Positive) refers to the number of instances correctly classified as belonging to the positive class, whereas, FP (False Positive) refers to the number of instances incorrectly classified to the positive class. In this context, recall is defined as the number of true positives divided by the total number of elements that actually belong to the positive class and denoted as the ratio: Recall = TP/(TP+FN), where, TP (True Positive) refers to the number of instances correctly classified as belonging to the positive class and FN (False Negative) refers to the number of positive instances incorrectly classified as belonging to the negative class. While there is a class imbalanced in the dataset, F-measure balances the performance of different classes in the classification model and is defined as the ratio of recall and precision: F-measure = 2 (Recall*Precision)/Recall + Precision.

6.3.2.4.1 Feature Selection

Feature selection is a task to select the minimum number of attributes/variables needed to accurately represent the data. By using relevant features, classification algorithms can improve their predictive accuracy, shorten the learning period and result in the simpler concepts. In selecting significant or relevant attribute, Ahmad and Dey's (2004) proposed "Significant Attribute Evaluator" method in WEKA was used. In this method, the significance of an attribute is evaluated by calculating the probabilistic significance as a two-way function of its association to the class decision which are attribute-to-class association and class-to-attribute association (Ahmad and Dey, 2004). An attribute is really significant if both attribute-to-class association and class-to-attribute association for the attribute are high. Finally, it provides a ranked list of the attributes where top ranked attributes are the most significant attributes than others.

#### 6.3.2.4.2 Decision Tree (DT)

Decision Tree (DT) is simple and widely used in classification and prediction. It is simple yet a powerful way of knowledge representation. Moreover, the classification tree models have some advantages over traditional statistical models (Kovacic, 2010). For example, classification trees can handle a large number of predictor variables; and classification tree models are non-parametric and can capture nonlinear relationships and complex interactions between predictors and dependent variables. In addition, decision trees are very popular because they produce classification rules that are easy to interpret than other classification methods (Yadav and Pal, 2012).

A decision tree is a flowchart in a tree-like structure, where each internal node is denoted by ovals, and leaf nodes are denoted by rectangles. All internal nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Arcs from an internal node to its children are labelled with distinct outcomes of the test. Each leaf node has a class label associated with it.

The decision tree has two phases (Han and Kamber, 2000):

- Growth phase or Build phase.
- Pruning phase.

To construct a tree the data is divided into two sets. One set is used to learn the tree and the other set is used to test the tree thereafter. For growing the tree, the first task is to find the root node for the tree. The root node is the first node splitting the entire dataset into two parts. The initial split at the root creates two new nodes, called branch nodes. The algorithm searches at both branch nodes again for the best split to separate the subsets. Following this recursive procedure, the algorithm continues to split all branch nodes by exhaustive search until either a branch node contains only patterns of one kind, or the diversity cannot be increased by splitting the node. The nodes where the tree is not further split, are labelled as leaf nodes. When the entire tree is split until only leaf nodes remain, the final tree is obtained (Breiman *et al.*, 1984; Quinlan, 1993). The tree may overfit the data.

Pruning is the process of reducing a tree by turning some branch nodes into leaf nodes, and removing the leaf nodes under the original branch. The pruning phase handles the problem of over fitting the data in the decision tree. The prune phase generalizes the tree by removing the noise and outliers. The accuracy of the classification increases in the pruning phase. Pruning phase accesses only the fully-grown tree. The growth phase requires multiple passes over the training data.

C4.5 Decision tree algorithm is developed by Quinlan Ross (Quinlan, 1993). C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. C4.5 uses gain ratio as an attribute selection measure to build a decision tree. Gain ratio removes the biasness of information gain when there are many outcome values of an attribute. At first, C4.5 Decision Tree algorithm calculates the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

DT is popular in predicting students' academic performance and has been used in many studies to predict students' performance/marks (Al-Radaideh *et al.*, 2006; Bharadwaj and Pal, 2011b; Yadav *et al.*, 2011; Yadev and Pal, 2012) to find out poor performing students so that institutions can arrange better support for them to succeed. In this experiment DT is used due to the advantages it provides in classification and also to be identical with its base model as the authors of the base model (Yadav and Pal, 2012) applied the same method in their study of students' performance prediction.

#### 6.3.2.4.3 10-fold Cross Validation

To evaluate the models, 10-fold cross validation method was used. In this method, data is split into 10 subsets of equal size set and then each subset in turn is used for testing and all the remaining sets are used as training set, and finally accuracy of each turn are averaged to estimate the overall accuracy rate. This is the most widely used cross-validation method.

## 6.4 Summary

This chapter presents the experimental methodology and platform used to test the hypotheses of this research study. Additionally, this chapter presents a simple ontology for student retention, progression and completion.

The following two chapters describe the outcome of the experiments described in this chapter. Chapter Seven presents the first experiment, which explored predictive models to identify at-risk students in their study using institutional internal datasets and external open data sources instead of traditional questionnaire. Chapter Eight presents the second experiment, which explored predictive models for student performance based on institutional internal datasets and external open data sources.

# Chapter 7: Exploring Student Progression Predictive Models based on Institutional Internal Datasets and External Open Data Sources

Research in student retention and progression to completion is traditionally survey-based, where researchers collect data through questionnaires and student interviews. The major issues with survey-based study are the potentially low response rates and the high cost associated with it. Nevertheless, a large number of datasets that could inform the questions that students are explicitly asked in surveys is commonly available in the external open datasets. This chapter describes the first experiment in detail, which explores a new way of developing student predictive models for student progression to completion that rely on the data available in institutional internal databases and external open data, without the need for surveys. The results of the empirical study for undergraduate students in their first year of study shows that predictive model based on institutional internal and external open data sources can perform as well as or even out-perform traditional survey-based ones.

## 7.1 Introduction

Student retention and progression to completion is one of the key issues to be addressed by higher education institutions around the world (Crosling *et al.*, 2009). Increasing student retention is a long-term goal in all academic institutions. The consequences of students dropping out are significant for students, academic staff, administrative staff and the institution itself. Since one of the criteria for government funding in tertiary education in the UK is the level of retention rate, both academic and administrative staff are under pressure to come up with strategies that could increase retention rates. The first year of study is recognized as a key stage, as during this period a new student is most likely to dropout from higher education institutions (Thomas *et al.*, 1996; Tinto, 1998; Yorke, 1999; Harvey *et al.*, 2006). Yorke (1999) noted about one third of students and Thomas *et. al* (1996) noticed about 77% of students withdraw from their courses during their first year. The indicators published by Higher Education Statistics Agency (HESA[1]), the rate of non-continuation rate in the UK higher education after one year of study varied from 7.9 to 9.5 between 2001/02 and 2009/10. The disproportionate number of students who leave higher education is a major problem and is the focus of studies on student retention. A number of theoretical models have been developed on student retention from many years. The first and most commonly used model is Tinto's model (1975; 1987; 1993), proposing a multivariate model of student retention in universities and colleges to explain early student departure; where the likelihood of a student withdrawing from higher education is seen as being determined by individual attributes, familial attributes, prior qualifications, social integration, academic integration, individual commitment, institutional commitment and external family and societal factors taking place during the course of study. Tinto claims that students who are highly integrated/engaged academically are more likely to continue and complete their degrees. These students have more friends at their university, have more personal contact with academics, enjoy being at the university, and thus are more likely to make the decision to remain in that environment. Research on factors related to student retention has traditionally relied on surveying a student cohort and following them for a specified period of time to determine whether they ultimately dropped out or continued their education. Using this design, researchers have worked to validate theoretical models of student retention including Tinto's widely employed model of student integration (Pascarella and Terenzini, 1980; Terenzini and Pascarella, 1980; Pascarella *et al.*, 1983).

Although it has been successfully used to-date, survey-based research may be too burdensome to sustain, as individual institutions may not have the capacity to construct and administer a similar instrument to study their unique retention situation. Even if an institution is capable of fielding a one-time

[1] http://www.hesa.ac.uk/index.php?option=com_content&task=view&id=2064&Itemid=1

retention survey, repeated administrations over time may be too overbearing. Moreover, another major limitation of survey-based test is low participation rates, which may often compromise the precision of the output. Thus, it is key for enrolment professionals and researchers to have sufficient means of evaluating the trends in the circumstances of student retention at their institution in order to develop or adjust support programs accordingly. Data-informed decision-making helps higher education institutions know whether they are achieving their missions (Schwartz *et al.*, 2010). Institutions routinely collect a broad array of information on their students' backgrounds and academic progress. Also, in the UK the Higher Education Statistics Agency (HESA[2]), the Higher Education Funding Council for England (HEFCE[3]), the Office for National Statistics (ONS[4]), and Unistats[5] routinely publish some open datasets. They can be used to develop student predictive models in the place of questionnaire-based predictive models that have been used to-date.

Although a large amount of data is available, the aggregation of data from institutional internal databases and external open data sources presents certain challenges. For example, data is in different locations with different formats and often with different identifiers. As discussed in Chapter 4, linked data approach has a strong impact on integrating and interlinking data of any kind. Linked data is interlinked RDF data that enables users to retrieve quality information from different data sources[6]. In this study, we examine the sufficiency of existing linked data standards and datasets in supporting student retention, progression and completion.

In section 7.2, we define the methodology of this experiment, in section 7.3, we explain the experiment and the attained results; and section 7.4 discusses the findings of the experiment, while the last section 7.5 presents the summary of this chapter.

[2] http://www.hesa.ac.uk/
[3] http://www.hefce.ac.uk/
[4] http://www.ons.gov.uk/ons/index.html
[5] http://unistats.direct.gov.uk/
[6] http://www.w3.org/DesignIssues/LinkedData.html

## 7.2 Methodology

The purpose of this experiment is to explore the potential of the new predictive models that rely on data commonly available in institutional internal databases and external open data sources instead of questionnaires used in the traditional student predictive models. We developed predictive models using the variable sets considered by Pascarella *et al.* (1983) in their study of first year student retention based on Tinto's theory of integration (1975). In their study, they used a set of questionnaires called Institutional Integration Scale (IIS) developed by Pascarella and Terenzini (1980) to measure various dimensions identified by Tinto as corresponding to the likelihood of persistence, which is being traditionally used in retention literature for many years. We subsequently developed predictive models that relied (i) only on information that is available in internal institutional databases and (ii) on information available in institutional internal databases and external open data sources. Using the predictive model by Pascarella *et al.* (1983) as a baseline we were able to explore the suitability of the proposed models.

### 7.2.1 Data and Data Sources

In this experiment, we consider 3 types of variables a) variables from institutional internal data sources (IDS), b) variables from traditional questionnaires/institutional integration scale (IIS) and c) variables from institutional external data sources/open data sources (EDS). Table 7.1 provides the list of all variables used in this experiment with their respective sources. In this experiment, National Student Survey (NSS) result published in Unistats website (as external/open data sources) is used to replace IIS variables, which measures student's academic and intellectual development, faculty student interaction and faculty concern for student development. Every year the NSS is conducted to measure students' satisfaction in different dimensions of their study subjects in their institutions such as satisfaction in teaching and learning, assessment and feedback, academic support, organization and management, learning resources and personal development. As IIS were also used to measure different dimensions of student satisfaction and integration (see in Table 7.2a), we consider a total 16 questionnaire items (see in Table 7.2b) from NSS which are related to student faculty interaction, faculty concern for student development, students' development and about their course among the 22 common questionnaires for all subjects as a replacement of the IIS questionnaire.

NSS measures students' satisfaction on their programme of study in a 5 points scale (Definitely Disagree, Moderately Disagree, Neither Agree nor Disagree, Moderately Agree, Definitely Agree). Unistats does not publish individual student data on the web, it publishes the percentages of respondents in each scale for an individual course. We have considered the actual value of the proportion of respondents who "agree" or "strongly agree" for those 16 questions for each individual course of the university of Southampton to include in this experiment. The published results for the academic year 2010-2011 are incorporated in this experiment.

Table 7.1 List of variables and variable sources for experiment 1.

| **Variable** | **Variable source** |
|---|---|
| Gender, Ethnicity, A Level tariff points, Accommodation Type, First year's first semester marks, Source of tuition fee, Study field | IDS |
| Parents' HE qualification | IDS |
| Student's working status in their first year of | Questionnaire item |
| Peer Group interaction (7 items/variables) | Questionnaire (IIS) |
| Student-Faculty interaction (5 items/variables) | Questionnaire (IIS) |
| Faculty Concern For Student Development and Teaching (5 items/variables) | Questionnaire (IIS) |
| Academic and Intellectual Development (7 items/variables) | Questionnaire (IIS) |
| Goal Commitment I | Questionnaire (IIS) |
| Institutional Commitment I | Questionnaire (IIS) |
| Goal Commitment II | Questionnaire (IIS) |
| Institutional Commitment II (2 items/variables) | Questionnaire (IIS) |
| Intention | Questionnaire (IIS) |
| The teaching on my course (4 items/variables) | EDS (Unistats) |
| Assessment and feedback (5 items/variables) | EDS (Unistats) |
| Academic support (3 items/variables) | EDS (Unistats) |
| Personal development (3 items/variables) | EDS (Unistats) |
| Overall satisfaction with the quality of the | EDS (Unistats) |

*IDS: Institutional Internal Data sources, EDS: Institutional External Data sources. A more detail description of these variables can be found in *Appendix E*.

Table 7.2a List of traditional questionnaire (IIS).

| **Traditional Questionnaire (IIS)** |
|---|
| Since coming to this university, I have made close personal relationship with other students. |
| The student friendships I have developed at the university have been personally satisfying. |
| My interpersonal relationships with other students have had a positive influence on my personal growth, attitudes, and values. |
| My interpersonal relationships with other students have had a positive influence on my intellectual growth and interest in ideas. |
| It has been difficult for me to meet and make friends with other students. |
| Few of the students I know would be willing to listen to me and help me if I had a personal problem. |
| Most students at this university have values and attitudes different from my own. |
| My non-classroom interactions with faculty have had a positive influence on my personal growth, values and attitudes. |
| My non-classroom interactions with faculty have had a positive influence on my intellectual growth and interest in ideas. |
| My non-classroom interactions with faculty have had a positive influence on my career goals and aspirations. |
| Since coming to this university, I have developed a close, personal relationship with at least one faculty member. |
| I am satisfied with the opportunities to meet and interact informally with faculty members. |
| Few of the faculty members I have had contact with are generally interested in students. |
| Few of the faculty members I have had contact with are generally outstanding and superior teachers. |
| Few of the faculty members I have had contact with are willing to spend time outside of class to discuss issues of interest and importance to students. |
| Most of the faculty members I have had contact with are interested in helping students grow in more than just academic areas. |
| Most faculty members I have had contact with are genuinely interested in teaching. |

Table 7.2a List of traditional questionnaire (IIS) (cont.).

| **Traditional Questionnaire (IIS)** |
|---|
| I am satisfied with the extent of my intellectual development since enrolling in this university. |
| My academic experience has had a positive influence on my intellectual growth and interest in ideas. |
| I am satisfied with my academic experience at this university. |
| Few of my courses this year have been intellectually stimulating. |
| My interest in ideas and intellectual matters has increased since coming to this university. |
| I am more likely to attend a cultural event now than I was before coming to this university. |
| Your choice of this institution was? |
| My academic performance has met my expectation. |
| It is important for me to graduate from this university. |
| I am confident that I have made the right decision in choosing to attend this university. |
| Getting good result is not important to me. |
| What is the highest expected academic degree? |
| It is likely that I will register at this university next year. |

Table 7.2b List of NSS questionnaires.

| **NSS questionnaires** |
|---|
| Staff are good at explaining things. |
| Staff have made the subject interesting. |
| Staff are enthusiastic about what they are teaching. |
| The course is intellectually stimulating. |
| The criteria used in marking have been clear in advance. |
| Assessment arrangements and marking have been fair. |
| Feedback on my work has been prompt. |
| I have received detailed comments on my work. |
| Feedback on my work has helped me clarify things I did not understand. |
| I have received sufficient advice and support with my studies. |
| I have been able to contact staff when I needed to. |
| Good advice was available when I needed to make study choices. |
| The course has helped me present myself with confidence. |
| My communication skills have improved. |
| As a result of the course, I feel confident in tackling unfamiliar problems. |
| Overall, I am satisfied with the quality of the course. |

### 7.2.2 Design of Empirical Study

For the purposes of this study, we asked the students to fill out questionnaires. Due to the limitation of database access permission, we asked students to provide us with both the information that is already available in internal university databases (such as the admission database and the student academic performance database) and for additional information that would normally be collected for the predictive models as per Pascarella *et al* (1983). In the first stage of this study, all students who enrolled in the academic year 2010/2011 were asked to complete an online questionnaire. We offered a small incentive to students to participate in our study (participation in a draw for vouchers with an online retailer) and obtained ethics approval from the university to conduct the questionnaire session. The total number of participants/respondents in this study was 149, of which about 15% are in the at-risk student group and 85% are in the not at-risk student group.

Apart from these data from institutional internal sources and questionnaire, we used NSS data to replace some traditional questionnaire data to develop and evaluate different student predictive models. The traditional questionnaire, the Institutional Integration Scale (IIS), which is traditionally used in retention study (Pascarella *et al.*, 1983; Herzog, 2005; Caison, 2007) for many years includes questionnaires about student's academic and intellectual development, academic support and satisfaction on teaching and learning. NSS also have some similar measurements of questionnaire.

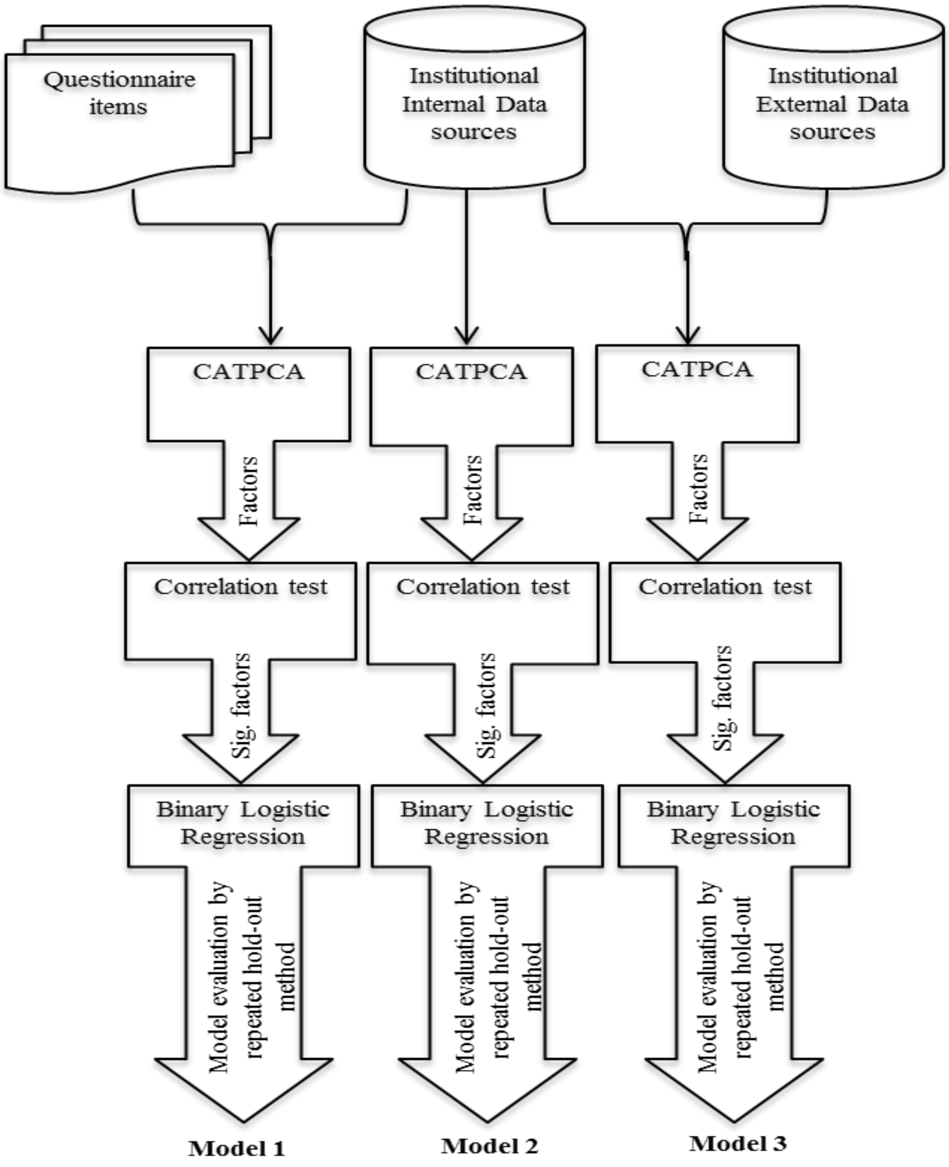

Questionnaire items
Institutional Internal Data sources
Institutional External Data sources
CATPCA
CATPCA
CATPCA
Factors
Factors
Factors
Correlation test
Correlation test
Correlation test
Sig. factors
Sig. factors
Sig. factors
Binary Logistic Regression
Binary Logistic Regression
Binary Logistic Regression
Model evaluation by repeated hold-out method
Model evaluation by repeated hold-out method
Model evaluation by repeated hold-out method
Model 1
Model 2
Model 3

Figure 7.1 Study framework to develop student progression models.

We explored whether we could replace those traditional questionnaires about student's academic and intellectual development, academic support and satisfaction on teaching and learning with the NSS questionnaires. Table 7.2 presents the IIS questionnaire items and 16 NSS questionnaires (considered as the replacement of the ISS questionnaire to include in the predictive models).

In this experiment, we developed three predictive models (model 1, model 2 and model 3), where the first model (model 1) includes all the independent variables considered by Pascarella *et al.* (1983) to develop the predictive model to find probable withdrawal students in their first year of study, which is a survey-based model. Second model (model 2) includes subset of variables from model 1, which are commonly available in the institutional internal databases to perceive how the model performs with only the available data in the institutions' databases. Finally, model 3 includes all the variables from model 2 (only institutional internal database variables) and includes new variables from external open data source as the replacement of the traditional questionnaire items/variables from model 1. Linked data technologies are used in integrating data from these different data sources. Figure 7.1 presents the study framework to develop these three different predictive models.

### 7.2.3 Data Analysis

The objective of data analysis was to establish:

- Whether it is possible to have a valid predictive model by omitting questions that are not available in institutional datasets

- Whether it is possible to have a valid and precise predictive model by replacing questions that would be asked in surveys by related data found in the open data cloud.

To achieve the objectives listed above, an exploration of the contribution of a number of variables to the predictive model was necessary. Categorical Principal Component Analysis (CATPCA) and logistic regression (LR) were used in this study. The goal of PCA is to reduce an original set of variables into a smaller set of uncorrelated components that represent most of the information found in the original variables. The technique is most useful when a large number of variables prohibit effective interpretation of the relationships between objects. By reducing the dimensionality, we interpret a few components rather than a large number of variables. CATPCA is an optimal scaling method belonging to the nonlinear multivariate analysis techniques. It is the nonlinear equivalent of PCA: it aims at the same goals of traditional PCA, but is suited for variables of mixed measurement level that may not be linearly related to each other (Linting *et al.*, 2007). The nonlinear PCA method is especially suitable for the dimension reduction problem with ordinal variables; because it simultaneously takes into account the nature of items, the different role of items in determining the measure, and the possible multidimensionality of the concept.

In this study, CATPCA was applied to overcome the multicollinearity problem and to reduce the model complexity. Multicollinearity is the situation where predictor variables are strongly correlated with each other and if it exists in the predictors, it can mislead the model output. Moreover, CATPCA was applied to extract factors (F) as well as to discover the factors/components structure, which are significantly correlated with the student outcome status. We followed Kaiser's rule to retain the factors for further analysis, if the analysis has more than 30 input variables, factors with eigenvalues greater than 1 are normally retained while it is recommended to retain factors (F) with eigenvalues greater than 0.7 with input variables less than 30 input variables (Field, 2009). Also the variable factor loadings which were smaller than 0.4 were ignored, that is if a variable's loading on a factor was found to be smaller than 0.4, it did not come towards the factor. To further optimize factor loadings, the varimax rotation algorithm with Kaiser normalization was applied to the resulting factor matrix. The varimax rotation is the most popular of all rotation algorithms. It aims to produce a few high valued loadings and many low-valued loadings so that the number of variables per factor is minimal with each variable having a maximum loading with regards to that factor (Abdi, 2003).

To enable further analysis with the data set using factors rather than variables, factor scores were saved in the data set using the Anderson-Rubin method as recommended by (Field, 2009). This method ensures that there are no correlations between factor scores. In the next step a correlation test was applied on retained factor scores and the students' outcome status (at-risk, not at-risk) looking for relationships between factors and students' outcome status.

Finally, logistic regression was applied to develop the predictive models with the significant factors only. Logistic regression is used when the dependent variables are categorical, rather than continuous. We used binary logistic regression, as our dependent variable has two categories (at-risk and not at-risk). The repeated hold-out method was applied to validate the predictive models. The cases (dataset) were randomly divided into two sets, where training set containing 70% of the cases and the test set containing the rest 30% of the cases. The training set was used to train the model and test set was used to validate the model. With this method, the predictive model can be made reliable by repeating the training and testing process through randomly partitioning the dataset, and average the accuracy rate of all repetition to produce the overall accuracy rate (Duda *et al.*, 2001).

Table 7.3 Correlation between 13 components and students' outcome variable for model 1 (survey-based model)

| Factor | **F1** | **F2** | **F3** | **F4** | **F5** | **F6** | **F7** | **F8** | **F9** | **F10** | **F11** | **F12** | **F13** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Outcome status | .006 | -.042 | -.023 | -.056 | **-.184*** | -.015 | .064 | **-.421**** | .033 | .106 | **.273**** | **-.207*** | -.125 |
| | .945 | .615 | .783 | .499 | **.025** | .852 | .439 | **.000** | .686 | .200 | **.001** | **.011** | .129 |

*Correlation is significant at the 0.05 level. ** Correlation is significant at the 0.01 level.

Table 7.4 Component structures of the significant components for model 1(survey-based model)

| **F5** | **F8** | **F11** | **F12** |
|---|---|---|---|
| Most of the faculty members I have had contact with are interested in helping students grow in more than just academic areas. | Intention. | My academic performance has met my expectation. | First Year $1^{st}$ Semester mark |
| I am satisfied with the opportunities to meet and interact informally with faculty members. | It is important for me to graduate from this university. | Parents' HE qualification | A level points |
| Most faculty members I have had contact with are genuinely interested in teaching. | | First Year $1^{st}$ Semester mark. | |
| Few of the faculty members I have had contact with are generally interested in students. | | | |
| My interest in ideas and intellectual matters has increased since coming to this university. | | | |

* The highest loading variables put first in the table and the lowest loading variables are in the last of the table.

Finally, the models were compared using the measures of overall prediction performance of the models, which derived from the confusion matrix that produced true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) outcome. As evaluating the model based on only the overall model accuracy (TP+TN)/(TP+TN+FP+FN) is not justified, because it is dominated by the student-in-good-standing class. Therefore, following the recommendation from Kotsiantis *et al.* (2004) and Lauria *et al.* (2012), two additional accuracy rate: sensitivity (TP/(TP+FN)) and specificity (TN/(TN+FP)) are also measured to evaluate the models. In addition, type I and type II error rate are calculated to evaluate the robustness of the models.

## 7.3 Experiment and Results

The purpose of this study was to explore a new approach of developing predictive model that relies on data commonly available in institutional internal and external data sources instead of questionnaires. In this study, we developed three predictive models (model 1, model 2 and model 3). The first model (model 1) is a survey-based model based on Tinto's integration model. The second model (model 2) is solely database-based model (includes only the subset of variables from model 1, which are commonly available in the institutional internal databases). The final model (model 3) includes all the variables from model 2 (only institutional internal database variables) and includes new variables from external data source as the replacement of the traditional questionnaire items/variables from model 1. CATPCA was applied to extract factors/components and also to discover the factors/components structure. The structures of the factors provide us the information about which questionnaires or variables are associated with which factors.

For the first model (survey-based model), a total of 39 variables were used in CATPCA. Following the approach stated in the data analysis section a total of 13 factors were retained for the survey-based model (model 1) where these 13 factors explained 70.23% of the total variance in the 39 items. A correlation test was applied between these 13 variables and the students' outcome status, and found only 4 factors (5, 8, 11 and 12) are significantly correlated with the students' outcome status. Table 7.3 shows the 13 factors and the correlation test result between these 13 variables and the students' outcome status. Also the factors, which are significantly, correlate with the students' outcome status are summarized with their associated input variables in Table 7.4. Factor 5 composed with five input variables, factor 8 and 12 composed with 2 input variables each, and factor 11 composed with 3 input variables (see in Table 7.4). The highest loading variables put first in the table. Student predictive model was developed with these four significant factors using binary logistic regression and the total accuracy of the model achieved 88.86% with attaining a sensitivity of 84.33% and specificity of 89.57%. In addition, type I error and type II error of the model are 10.43% and 15.67% respectively. Furthermore, SPSS provides Cox and Snell's R Square and Nagelkerke's R Square values to explain the proportion of variation in the outcome variable explained by the model. Cox and Snell's R Square has the disadvantage that it may not achieve the maximum value of one, even when the model predicts all the outcomes perfectly. Nagelkerke's R Square is an improvement over Cox and Snell's R Square that can attain a value of one when the model predicts the data perfectly. For model 1, Cox and Snell's R Square and Nagelkerke's R Square achieved 0.27 and 0.54 respectively. Therefore, considering Nagelkerke's R Square, it can be said that model 1 explained about 54% of the variation in the data.

Table 7.5 Correlation between 5 components and students' outcome variable for model 2 (solely database based model)

| | F1 | F2 | F3 | F4 | F5 |
|---|---|---|---|---|---|
| Outcome status | .094<br>.256 | .123<br>.136 | **-.391****<br>**.000** | .067<br>.417 | **-.253****<br>**.002** |

** Correlation is significant at the 0.01 level.

Table 7.6 Component structures of the significant components for model 2 (solely database based model)

| F3 | F5 |
|---|---|
| A level points | Parents' HE qualification |
| First Year 1st Semester mark | |

* The highest loading variables put first in the table and the lowest loading variables are in the last of the table.

Utilising the same procedure we develop the second model (solely institutional internal database based model). We found only 8 variables are available in the institutional internal database among all of the 39 variables in model 1 (survey-based model). A total of five factors were retained after applying CATPCA that explained 72.47% of the total variance in the 8 database items for model 2 (solely institutional internal database based model) and two of them were significantly correlated with the students' outcome status, which is presented in Table 7.5. Table 7.6 presents the factors/components structure of these two significant factors. The total accuracy for solely institutional internal database based model (model 2) was achieved 84.94% where model sensitivity is 61.33%, specificity is 87.46%, type I error is 12.54% and type II error is 38.67%. Model 2 achieved Cox and Snell's R Square and Nagelkerke's R Square values 0.24 and 0.50 respectively. Therefore, this model explained about 50% of the variance in the dependent variable.

The third model (model 3) includes only the database items from model 1 (survey-based model) as well as data from external open data source to replace questionnaire data from model 1 (survey-based model). Total 24 input variables were considered to develop model 3 (8 database items and 16 NSS questionnaire items). The same method applied to develop model 3 (institutional internal database and external open dataset based model) as the previous two models. A total of eight factors were retained with explaining 88.77% of the total variance in the 24 items and among these eight factors three factors were found significantly correlated with the student output status. Table 7.7 presents the correlation results, while Table 7.8 presents the components/factors structure of the three significant factors. The total accuracy of model 3 (institutional internal database and external open dataset based model) was achieved 89.20% with model sensitivity 89.33% and specificity 89.63%. In addition, type I error and type II error were 10.37% and 10.67% respectively for this model. This model explained about 59% of the variance in the dependent variable, where, Cox and Snell's R Square and Nagelkerke's R Square achieved 0.29 and 0.59 respectively.

Table 7.7 Correlation between 8 components and students' outcome variable for model 3 (institutional internal database and external open dataset based model).

| | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 |
|---|---|---|---|---|---|---|---|---|
| Outcome status | -.028<br>.731 | **-.161***<br>**.050** | .065<br>.432 | .063<br>.447 | -.062<br>.454 | **-.377****<br>**.000** | .026<br>.754 | **-.284****<br>**.000** |

*Correlation is significant at the 0.05 level. ** Correlation is significant at the 0.01 level.

Table 7.8 Component structures of the significant components for model 3 (institutional internal database and external open dataset based model).

| F2 | F6 | F8 |
|---|---|---|
| Staff have made the subject interesting. | A level points | Parents' HE qualification |
| Staff are enthusiastic about what they are teaching. | First Year 1st Semester mark | |
| Field of Study | | |
| Gender | | |
| I have received detailed comments on my work. | | |

*The highest loading variables put first in the table and the lowest loading variables are in the last of the table.

Table 7.9 Summary of the three student predictive models.

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| Number of Input variables | 39 | 8 | 24 |
| Number of Factors retained after CATPCA | 13 | 5 | 8 |
| Number of significant factors | 4 | 2 | 3 |
| Sensitivity of the model (%) | 84.33 | 61.33 | 89.33 |
| Specificity of the model (%) | 89.57 | 87.46 | 89.63 |
| Type I error (%) | 10.43 | 12.54 | 10.37 |
| Type II error (%) | 15.67 | 38.67 | 10.67 |
| Cox and Snell's R Square | 0.27 | 0.24 | 0.29 |
| Nagelkerke's R Square | 0.54 | 0.50 | 0.59 |
| Total model accuracy (%) | 88.86 | 84.94 | 89.20 |

Where,

- Overall model accuracy = {(TP+TN)/(TP+TN+FP+FN)} *100
- Sensitivity = {TP/(TP+FN)} * 100
- Specificity = {TN/(TN+FP)} * 100
- Type I error = 1- specificity = {FP/(FP+TN)} * 100
- Type II error = 1- sensitivity = {FN/(TP+FN)} * 100

## 7.4 Discussion

Research on retention typically relies on surveying of student perceptions in relation to the factors believed to theoretically influence persistence decisions. However, this resource-intensive methodology is not always feasible for retention research at individual institutions. Caison (2007) compares traditional survey-based retention research methodology with an analytic approach that relies on data commonly available in institutional internal databases. His study result confirms that only the variables available from the institutional internal databases are not sufficient to build a good performing predictive model. It requires more additional information/data to perform better. Also, the same stands for the predictive model, which was developed using Pascarella and Terenzini's Institutional Integration Scale (IIS).

The current study result ratifies the above findings as the prediction model based on solely institutional internal databases (model 2) performed the lowest among the three predictive models. It achieved an overall model accuracy of 84.94%. When adding traditional questionnaires (IIS) with institutional internal databases (model 1), the prediction performance improved and achieved an overall model accuracy of 88.86%. However, when replacing the traditional questionnaire data using external open data sources, the prediction model (model 3) performed the best, attaining an overall model accuracy of 89.20%. Moreover, according to model sensitivity, the ability of the model to detect the student population who are truly at-risk in their programme of study, model 1, model 2 and model 3 achieved 84.33%, 61.33% and 89.33% respectively. Therefore, it reflects that type II error/false negative error rate for model 1, model 2 and model 3 are 15.67%, 38.67% and 10.67% respectively. Based on the sensitivity, it is observed that model 3 (based on institutional internal databases and external open data sources) provided the highest performance. In other words, this model attained the lowest proportion of type II error compared to the other two models. From the sensitivity (61.33%) of model 2 (solely based on institutional internal databases), it can be speculated that only information/data from institutional internal databases is not sufficient to predict students at-risk in their study.

In terms of specificity, model 1 (survey-based) attained 89.57%, model 2 (solely based on institutional internal databases) attained 87.46% and model 3 (institutional internal databases and external open dataset based model) attained 89.63% of the model specificity. This states that model 1, model 2 and model 3 produced respectively 10.43%, 12.54% and 10.37% false positive error/type I error. Similar to type II error, model 3 (institutional internal databases and external open dataset based model) provided the lowest proportion of type I error among the three predictive models. It can be noticed that the specificity of model 1 (survey-based) and model 3 (institutional internal databases and external open dataset based model) are comparable. Moreover, model 2 achieved much higher specificity (87.46%) compared to its corresponding sensitivity (61.33%).

The consequence of misclassification of at-risk students as not at-risk (type II error) is that these students would not receive additional learning support provided to the students at-risk because they will be classified among not at-risk students due to the lack of robustness of the model. On the other hand, the consequence of misclassification of not at-risk students among at-risk students (type I error) is that HEI need to arrange additional learning support for these students even though they do not need it. Therefore, it is desirable to have both error rates smaller in a good performing predictive model.

Based on the component structures of the three predictive models, this study strongly supports that students' A levels point (pre-entry qualifications) and current academic marks are the strongest determinant to identify at-risk students in their first year of study. Also, the correlation test of the three predictive models established that students with high A level points and first semester marks are less likely to be at-risk in their programme of study. Accordingly, it is speculated that these students are more serious about completion of their study. Similarly, parents' HE qualifications were found to be important in detecting at-risk students according to the component structures of all three models. Also, the results of the correlation tests for the models confirm that students whose parents have higher education qualifications are less likely to be at-risk in their programme of study. This implies that students with highly educated parents are more aware of the importance of higher education and encourage their children to get university degrees. Moreover, the component structure of model 3 indicated that, in addition to students' A level points, first semester marks and parents' higher education qualifications; 3 NSS questionnaires (i.e., "Staff have made the subject interesting", "Staff are enthusiastic about what they are teaching" and "I have received detailed comments on my work") were significant for identifying at-risk students in their first year of study.

It is noteworthy that all the three models explained relatively good percent of variance in the dependant variable. Model 1 explained 54 percent, model 2 explained 50 percent and model 3 explained 59 percent of the variance in the dependant variable. The proportions achieved in the current study are quite comparable to some of the previous studies. For example, Allen and Nelson (1989) achieved from 44% to 53%, Milem and Berger (1997) achieved 41% , Berger (1997) achieved 42%, Berger and Braxton (1998) achieved 44% of the variance in retention.

The results of this study strongly support the use of institutional internal databases and external open data sources to conduct institution specific retention and progression to completion research in order to identify at-risk students and arrange intervention programs for them. The findings of this study do not weaken the results of the model developed using traditional questionnaires; rather, this study offers researchers a new approach to engage in retention studies. This expanded toolkit for retention research offers the possibility for more research in diverse settings which, given resource constraints, would not have otherwise been possible. This study lays the groundwork for this effort.

However, the current study has several limitations. Firstly, the current study contained an unbalanced dataset, i.e. the not at-risk student group has a larger number of trials than the at-risk student group, which may bias the study results. Though we considered additional measures (sensitivity, specificity, type I error rate and type II error rate) along with overall model accuracy to evaluate the predictive models, we recommend that future studies include an adequate number of trials in both student groups, preferably a balanced dataset. Secondly, we could not classify the real dropout students because we did not have means to contact them and we did not have access or permission to the University of Southampton's student database to collect their information. It is expected that in the future higher education institutions would allow access to information of the dropout students. Consequently, integration of such datasets will support the development of a more robust model. Finally, all the students' specific data provided by students themselves during the online questionnaire session, which may leads, some wrong information and students self reported data may influence model output. Nonetheless, in the future higher education institutions could allow access to detailed information of the students and this may reduce the dependency to use self-reported questionnaires.

## 7.5 Summary

This chapter presents experiment 1, which explores a new method of developing student predictive models through integrating data from institutional internal databases and external open data sources without having to rely on questionnaires that have been essential in existing predictive models. Moreover, this experiment tests hypotheses 1, 2 and 3 regarding the sufficiency of linked data technologies and external open data sources in developing student predictive model to support student retention, progression and completion. This experiment applied categorical principal component analysis and logistic regression to develop student predictive models and used repeated hold-out method to evaluate the robustness of the predictive models. In addition, linked data technologies are used to integrate data from various data sources.

The results of the experiment provide evidence that the model based on institutional internal databases and external open data sources performs better than other two predictive models (survey-based model and model solely based upon institutional internal databases). The survey-based model performs second best while the lowest performing model was solely institutional internal database based predictive model.

This study supports the prospects of external open data sources in developing student predictive models to support student retention and progression to completion instead of questionnaires. Moreover, this study evidences the prospects of linked data technologies in institutional research to support student retention and progression to completion as the potential data are spread out in different institutional internal and external open data sources. We believe that the results of this experiment may increasingly improve the design of future students' predictive models to support student retention. The next chapter presents experiment 2, which explores student predictive model to predict students' first year final marks using institutional internal databases and external open data sources.

# Chapter 8: Students' Performance Prediction using Institutional Internal Datasets and External Open Data Sources

The methodologies of the experiments conducted in this research are described in Chapter 6. The experimental design and the data analysis results of the first experiment have been presented in Chapter 7. This chapter presents the design and the data analysis results of the second experiment. In this experiment, undergraduate student's mark prediction models have been developed using institutional internal databases and external open data sources. The results of the experiment for undergraduate students' first year mark prediction show that prediction based on institutional internal databases and external open data sources can provide improved prediction compared to the model based on only institutional internal data sources.

## 8.1 Introduction

The main objective of higher education institutions is to provide quality education and to improve student success. Efficient prediction of student performance in higher education institutions is one way to reach the highest level of quality in the higher education system. Timely intervention, based on early identification of poor performance, is likely to help weaker students improve their performance so that they can progress and complete their study successfully. For example, if educational institutions can predict students' academic performance early before their examination, then extra effort can be taken to arrange proper support for the lower performing students to improve their studies and help them to succeed.

The prediction of student performance with high accuracy is beneficial to identify the students with low academic achievements initially. It is required that the teacher assist the identified students more so that their performance is improved in the future. Researchers used various classification methods in their studies to predict students' academic performance, such as decision trees, classification and regression trees, logistic regression, bayesian classification, support vector machine and neural network (Kotsiantis *et al.,* 2004; Al-Radaideh *et al.,* 2006; Vandamme *et al.,* 2007; Kovacic, 2010; Yadev et al., 2011; Yadev and Pal, 2012). Among these, decision trees remain popular in predicting students' performance (Al-Radaideh *et al.*, 2006; Bharadwaj and Pal, 2011b; Yadav *et al.*, 2011; Yadev and Pal, 2012). A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. The decision tree starts with a root node on which users take actions. From this node, users split each node recursively according to the decision tree learning algorithm (e.g. ID3, C4.5 etc.). The final result is a tree in which each branch represents a possible scenario of decisions and its outcome. Among the decision tree algorithms, C4.5 is popular for its higher performance in classification accuracy (Yadav *et al.*, 2011; Yadev and Pal, 2012).

In Chapter 3, while discussing students' performance/marks prediction, we found that students' academic performance is based upon diverse factors like personal, social, psychological and other environmental variables. Various experiments have been carried out in this area to predict students' academic performance using institutional internal datasets and/or through conducting student questionnaires. Nowadays, a large amount of educational data is available from external open data sources. Although a large amount of data is available, the combination of data from different sources presents certain challenges (Arnold, 2010). For example, data is maintained in different locations, in different formats and often with different identifiers. To this end, linked data technologies are considered to be well suited for data integration and interoperability.

The aim of this experiment is to test hypothesis 4 regarding the sufficiency of existing external open data sources to predict students' first year mark. This experiment applied a decision tree algorithm to predict students' first year mark using institutional internal and external open data sources and used linked data technologies to integrate data from these disparate data sources.

Section 8.2 describes the experimental methodology, while section 8.3 provides the experiment results and discussion of the findings of the experiment. Finally, section 8.4 provides the summary of this chapter.

## 8.2 Methodology

The main goal of this experiment is to investigate how accurate or improved prediction models can be developed to predict students' first year marks/performance using institutional internal datasets and data commonly available in the external open data sources. This experiment considered institutional internal variables, which are commonly available in the institutional internal databases and external variables, which can be derived or can be integrated from external open data sources. This experiment incorporated the variables (as many as available) used by Yadav and Pal (2012) in their studies of predicting students' academic performance, as most of those variables are commonly available in the students' enrolment database. Yadav and Pal (2012) conducted their study with a classification tree to predict student academic performance using students' gender, admission type, previous school marks, medium of teaching, location of living, accommodation type, father's qualification, mother's qualification, father's occupation, mother's occupation, family annual income and so on. The results found in their study are significant. Using these variables, they obtained a high classification accuracy of 67.78% using the C4.5 classification technique.

In the first step, we developed two models, model 1 and model 2. Model 1 is developed using only the institutional internal variables and model 2 is developed using the institutional internal variables and external variables. Using the predictive model by Yadav and Pal (2012), we were able to explore the suitability of the external open data sources in predicting students' mark. Subsequently, we extended the above two predictive models as model 3 and model 4 by adding students' first semester mark to observe the effect of adding current (first semester marks) academic performance on the prediction performance of the models. Moreover, this will help us analyse the effect of external open data sources on both predictive models before and after adding current academic performance (first semester mark). In many studies current or previous academic performance/marks were significant in predicting students' final marks/performance (Kam and Ch'ng, 2009; Kovacic and Green, 2010). This experiment used WEKA[73] for data analysis. WEKA is an open source software that implements a large collection of machine learning algorithms and is widely used in data mining applications (Al-Radaideh *et al.*, 2006; Yadev and Pal, 2012). In this experiment, the decision tree classification technique has been applied in building the predictive models. We used the J48 decision tree algorithm to develop the mark prediction models. The J48 algorithm is the WEKA implementation of the C4.5 top-down decision tree learner proposed by Quinlan (1993). The same classification technique was applied by Yadav and Pal (2012) in their study of predicting academic performance and achieved the best classification accuracy. Al-Radaideh *et al.* (2006) and Yadev *et al.* (2011) also found the best classification accuracies while using the C4.5 decision tree algorithm in their studies. The 10-fold cross validation method was used to validate the model performance.

Finally, model 1 was compared to model 2 and model 3 was compared to model 4 based on accuracy, precision and recall of the models as recommended in Sehgal *et al.* (2012) to determine the best performing model among them. In addition, F-measure values of the models are considered to determine the best performing model, following the recommendation by Moore *et al.* (2009), who suggested that F-measure balances the performance of the classes when there is substantial class imbalance in a dataset. Furthermore, a Student's t-test was performed to confirm whether the models performances differ significantly or not. Figure 8.1 shows the study framework to develop these four predictive models to predict students' first year final marks.

### 8.2.1 Data and Data Sources

In this study, we considered two types of variables: a) variables from institutional internal data sources (IDS) and b) variables from institutional external (open) data sources (EDS).

[73] http://www.cs.waikato.ac.nz/ml/weka/

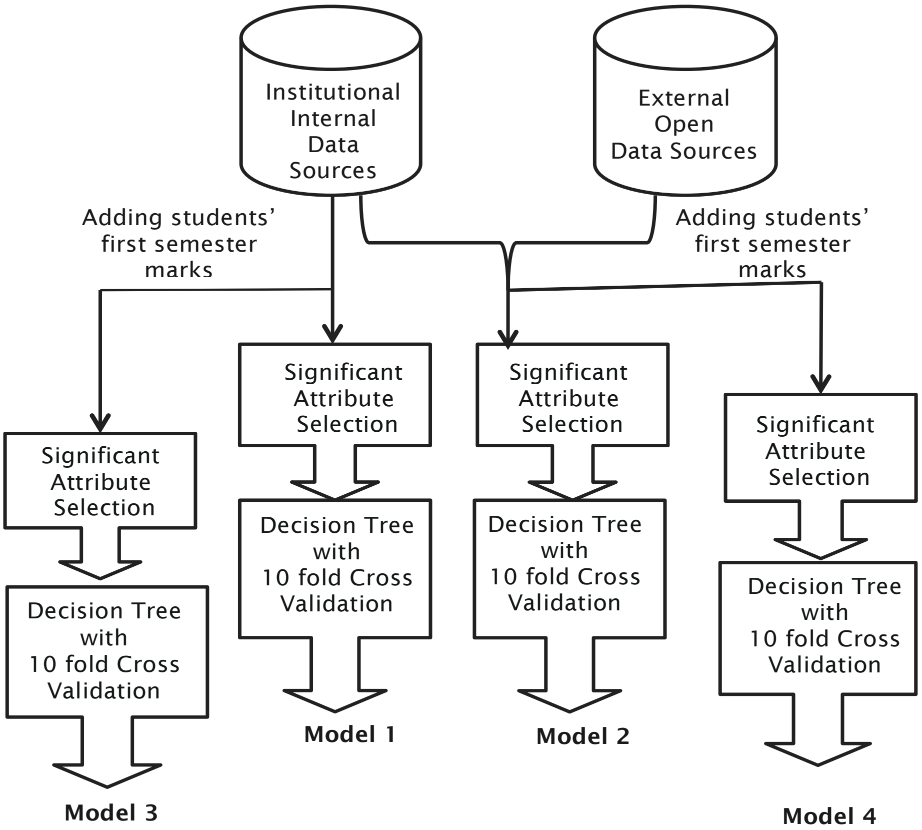

Institutional Internal Data Sources
External Open Data Sources
Adding students' first semester marks
Adding students' first semester marks
Significant Attribute Selection
Significant Attribute Selection
Significant Attribute Selection
Significant Attribute Selection
Decision Tree with 10 fold Cross Validation
Decision Tree with 10 fold Cross Validation
Decision Tree with 10 fold Cross Validation
Decision Tree with 10 fold Cross Validation
Model 1
Model 2
Model 3
Model 4

Figure 8.1 Study framework to develop students' first year marks prediction models.

In our previous study on identification of at-risk students based on institutional internal and external open data sources (see Chapter 7), we found that a predictive model using the National Student Survey (NSS) results performed better (or quite similar) than the model based on a traditional questionnaire and institutional internal data sources (survey-based model). We were motivated to include the same 16 NSS questions in this study as the external open data to predict students' first year final mark. NSS measures students' satisfaction in their programme of study on a 5 points scale (Definitely Disagree, Moderately Disagree, Neither Agree nor Disagree, Moderately Agree, Definitely Agree). The website publishes the percentages of respondents in each scale for an individual course. We have considered the actual value of the proportion of respondents who "agree" or "strongly agree" for those 16 questions for each individual course of the University of Southampton. We included the published dataset for the 2010-2011 academic year. Also, the Office for National Statistics' (ONS[74]) published data (gross annual income

[74] http://www.ons.gov.uk/ons/index.html

based on Standard Occupational Classification 2010, (SOC2010)) has been used in this study to derive parents' annual income and students' socio economic class. Moreover, we derived participants' neighbourhood group using the Higher Education Funding Council for England (HEFCE[75]) published open dataset. All other data for this experiment was collected by administering the same online questionnaire used for the previous experiment (experiment 1). The participants were all first-degree/undergraduate students who enrolled in the 2010-2011 academic year in any programme of study at the University of Southampton in the UK. Participants were recruited by a group email invitation as well as by circulating posters in the university campus area. Participants were required to answer the online questionnaires by following the provided link in the invitation. The total number of participants was 149, of which 32.89%, 43.62%, 19.46% and 4.02% of the students were in 71%-100%, 61%-70%, 51%-60% and 41%-50% first year final marks group, respectively. Table 8.1 provides the list of all the variables used in this experiment with their description, possible values and data sources.

### 8.2.2 Experiment

The objectives of this experiment are to

- examine the capability of institutional external open data sources to predict students' mark while combining with only students' enrolment data, and
- examine the capability of institutional external data sources while combining with students' enrolment data and current academic performance (students' first semester mark).

[75] http://www.hefce.ac.uk/

Table 8.1 List of variables with their description and sources for experiment 2.

| Variable name | Description and possible values | Variable Source |
|---|---|---|
| Study_field | Students field of study: Applied (engineering, physics, chemistry etc.), Non-applied (Languages etc.) | IDS |
| Gender | Students' gender/sex: Male, Female | IDS |
| Residence | Students Residence/Domicile: UK, Other-EU, Non-EU | IDS |
| A_level_point | Students result in A level or any other equivalent entry qualifications.<br>A*=140, A=120, B= 100, C=80, D=60<br>For example, if a student's A level grade is AAA then his A level point counted as AAA=120+120+120=360. | IDS |
| AdmissionType | Students' admission type: Direct, Clearing | IDS |
| Accom_Type | Students' accommodation type: University halls, Others | IDS |
| Parents_HE | Parents' higher education qualification: Yes, No | IDS |
| M_Occu_cat | Mother's occupation: Service, House-wife, NA | IDS |
| F_Occu_cat | Father's occupation: Service, Business, NA | IDS |
| FirstYr_1stSem_mark | Percentage of marks in first year's first semester: 71%-100%, 61%-70%, 51%-60%, 41%-50% | IDS |
| FirstYrMarkrange | Percentage of final marks in first year: 71%-100%, 61%-70%, 51%-60%, 41%-50% | IDS |
| Part_neighborhood | Students categorized according to their postcode: Lower participation neighborhood, Other neighborhood, Unknown | EDS (HEFCE) |
| ONS_soc_eco_class | Students' socio economic class based on parents' occupations: MP-occupations, I-occupations, RM-occupations | EDS (ONS) |
| P_annual_income | Parents' annual income. | EDS (ONS) |

Table 8.1 List of variables with their description and sources for experiment 2 (cont.).

| Variables name | Description and possible values | Variable Source |
|---|---|---|
| NSS_Q1 | Staffs are good at explaining things. | EDS (Unistats) |
| NSS_Q2 | Staffs have made the subject interesting. | EDS (Unistats) |
| NSS_Q3 | Staffs are enthusiastic about what they are teaching. | EDS (Unistats) |
| NSS_Q4 | The course is intellectually stimulating. | EDS (Unistats) |
| NSS_Q5 | The criteria used in marking have been clear in advance. | EDS (Unistats) |
| NSS_Q6 | Assessment arrangements and marking have been fair. | EDS (Unistats) |
| NSS_Q7 | Feedback on my work has been prompt. | EDS (Unistats) |
| NSS_Q8 | I have received detailed comments on my work. | EDS (Unistats) |
| NSS_Q9 | Feedback on my work has helped me clarify things I did not understand. | EDS (Unistats) |
| NSS_Q10 | I have received sufficient advice and support with my studies. | EDS (Unistats) |
| NSS_Q11 | I have been able to contact staff when I needed to. | EDS (Unistats) |
| NSS_Q12 | Good advice was available when I needed to make study choices. | EDS (Unistats) |
| NSS_Q19 | The course has helped me present myself with confidence. | EDS (Unistats) |
| NSS_Q20 | My communication skills have improved. | EDS (Unistats) |
| NSS_Q21 | As a result of the course, I feel confident in tackling unfamiliar problems. | EDS (Unistats) |
| NSS_Q22 | Overall, I am satisfied with the quality of the course. | EDS (Unistats) |

Therefore, for the above objectives, an analysis of the importance of the variables of the predictive model was necessary. We used Ahmad and Dey's (2004) proposed significant attribute evaluator method in WEKA to select the significant variables/attributes. Zhao and Luan (2006) suggested that even though some variables may have little significance to the overall prediction outcome, they can be essential to a specific record in the data set. Following this advice, we considered all the variables/attributes with a score value greater than "0" to develop the four predictive models as described in the methodology. The first model (model 1) is developed using only the institutional internal variables (without including students' first semester marks) and the second model (model 2) is developed using the institutional internal variables and external variables (without including students' first semester marks). We developed model 3 and model 4 by adding students' first semester marks with the previous two predictive models. Finally, to examine whether the models using external variables can obtain a similar or improved accuracy as traditional predictive models, a comparison was made between model 1 and model 2 and between model 3 and model 4 based on accuracy, precision recall and F-measure of the models. Moreover, a Student's t-test was performed using SPSS to ratify the significance of performances improvement.

## 8.3 Results and Discussion

For the first model (based on only institutional internal databases without students' first semester marks), we considered 9 input variables and found all of them to be significant. They scored greater than "0" for the prediction of the students' first year final marks. Table 8.2 provides the list of these ranked attributes/variables with their score. The highest scored attributes are more significant compared to the other variables/attributes. From Table 8.2, it is found that students' A level points is the top ranked variable and therefore, contribute most to predict students' first year final mark. After this variable, mother's occupation, field of study, admission type and father's occupation are ranked in order according to their scores.

For model 2 (based on institutional internal databases and external open dataset without students' first semester marks), we considered a total of 28 institutional internal and external variables, of which only 15 variables are found to be significant and scored greater than "0". Table 8.3 presents the list of these 15 attributes with their score value. Similar to model 1, it is observed that students' A level points is the most significant variable to predict students' first year final mark. Subsequently, NSS's five questionnaires (NSS_Q2, NSS_Q6, NSS_Q9, NSS_Q5 and NSS_Q8) are ranked as the second, third, fourth, fifth and sixth contributing variables to predict students' marks. Among other variables, mother's occupation, study field, admission type, father's occupation and socio-economic status are found to be significant.

For model 3 (based on institutional internal databases with students' first semester marks), all 10 input variables are found to be significant with a score value greater than "0". They are considered to be included in the prediction model. Table 8.4 provides the significant variables list with their score values for model 3. It can be noted that students' first semester mark is ranked as the top most significant variable. Subsequently, students' A level points, mother's occupation, study field, admission type and father's occupation are ranked to predict students' marks in their first year of study.

For model 4 (based on institutional internal databases and external open dataset with students' first semester marks), 17 out of 29 institutional internal and external variables are detected to be significant with a score value greater than "0" and therefore, these variables are selected to be included in the model development. Table 8.5 presents 17 significant variables and their respective scores for model 4. This model behaved similar to model 3 in that students' first semester mark is the top ranked significant variable to predict students' first year final mark. The second most significant variable is students' A level points, which is also similar to model 3. Like model 3, the same NSS's five questionnaires (NSS_Q2, NSS_Q6, NSS_Q8, NSS_Q5, and NSS_Q9) are selected and ranked as the third, fourth, fifth, sixth and seventh significant variables. After these, mother's occupation, study field, admission type, father's occupation and students' socio-economic status are ranked subsequently.

Table 8.2 Selected variables/attributes with their score for model 1 (institutional internal database based model without first semester marks)

| **Variable/Attribute name** | **Score** |
|---|---|
| A_level_point | 0.455 |
| M_Occu_cat | 0.262 |
| Study_field | 0.253 |
| AdmissionType | 0.225 |
| F_Occu_cat | 0.218 |
| Residence | 0.184 |
| Gender | 0.17 |
| P_HE | 0.126 |
| Accom_Type | 0.119 |

Table 8.3 Selected variables/attributes with their score for model 2 (institutional internal database and external open dataset based model without first semester marks)

| **Variable/Attribute name** | **Score** |
|---|---|
| A_level_point | 0.455 |
| NSS_Q2 | 0.335 |
| NSS_Q6 | 0.335 |
| NSS_Q9 | 0.325 |
| NSS_Q5 | 0.325 |
| NSS_Q8 | 0.325 |
| M_Occu_cat | 0.262 |
| Study_field | 0.253 |
| AdmissionType | 0.225 |
| F_Occu_cat | 0.218 |
| ONS_soc_eco_gp | 0.21 |
| Residence | 0.184 |
| Gender | 0.17 |
| P_HE | 0.126 |
| Accom_Type | 0.119 |

Table 8.4 Selected variables/attributes with their score for model 3 (institutional internal database based model with first semester marks)

| Variable/Attribute name | Score |
|---|---|
| FirstYr-1stSem_markrange | 0.781 |
| A_level_point | 0.455 |
| M_Occu_cat | 0.262 |
| Study_field | 0.253 |
| AdmissionType | 0.225 |
| F_Occu_cat | 0.218 |
| Residence | 0.184 |
| Gender | 0.17 |
| P_HE | 0.126 |
| Accom_Type | 0.119 |

Table 8.5 Selected variables/attributes with their score for model 4 (institutional internal database and external open dataset based model with first semester marks)

| Variable/Attribute name | Score |
|---|---|
| FirstYr-1stSem_markrange | 0.781 |
| A_level_point | 0.455 |
| NSS_Q2 | 0.335 |
| NSS_Q6 | 0.335 |
| NSS_Q8 | 0.325 |
| NSS_Q5 | 0.325 |
| NSS_Q9 | 0.325 |
| M_Occu_cat | 0.262 |
| Study_field | 0.253 |
| AdmissionType | 0.225 |
| F_Occu_cat | 0.218 |
| ONS_soc_eco_gp | 0.21 |
| Part_neighborhood | 0.186 |
| Residence | 0.184 |
| Gender | 0.17 |
| P_HE | 0.126 |
| Accom_Type | 0.119 |

Figures 8.1, 8.2, 8.3 and 8.4 present the classification rule generated by the J48 decision tree algorithm for model 1 (institutional internal database based model without students' first semester marks), model 2 (institutional internal database and external open data sources based model without students' first semester marks), model 3 (institutional internal database based model with first semester marks) and model 4 (institutional internal database and external open data sources based model with students' first semester marks) respectively.

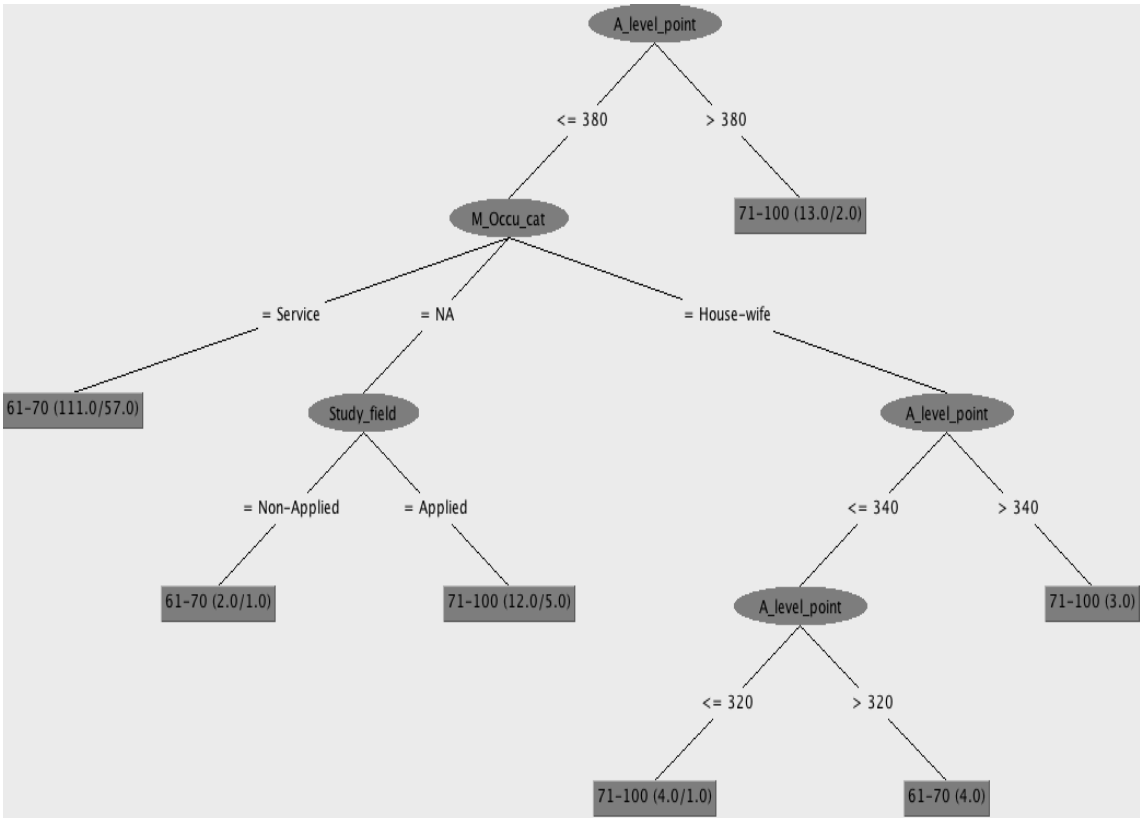


Figure 8.1 J48 rule for model 1(institutional internal database based model without students' first semester marks).

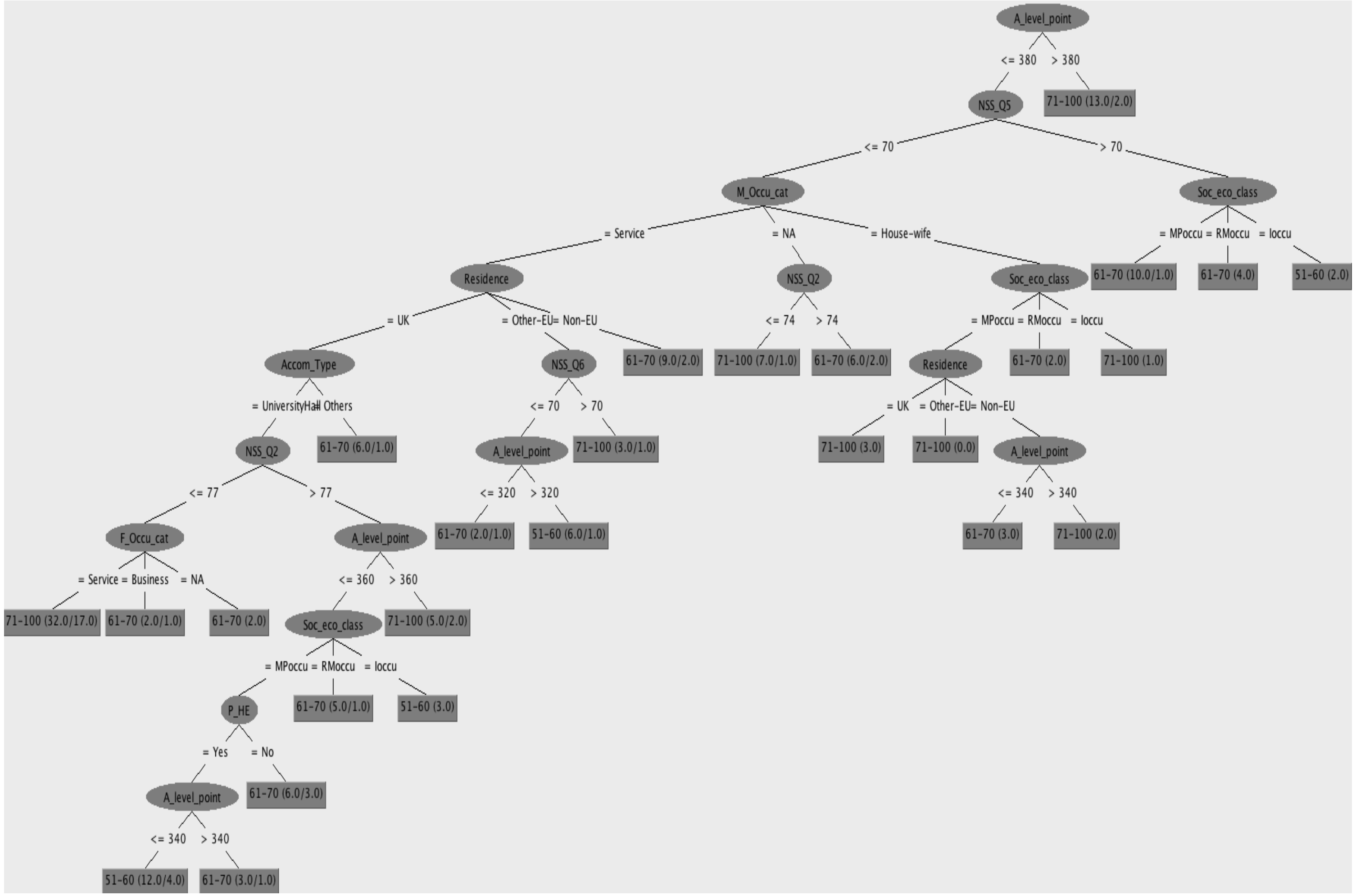


Figure 8.2 J48 rule for model 2 (institutional internal database and external open dataset based model without students' first semester marks).

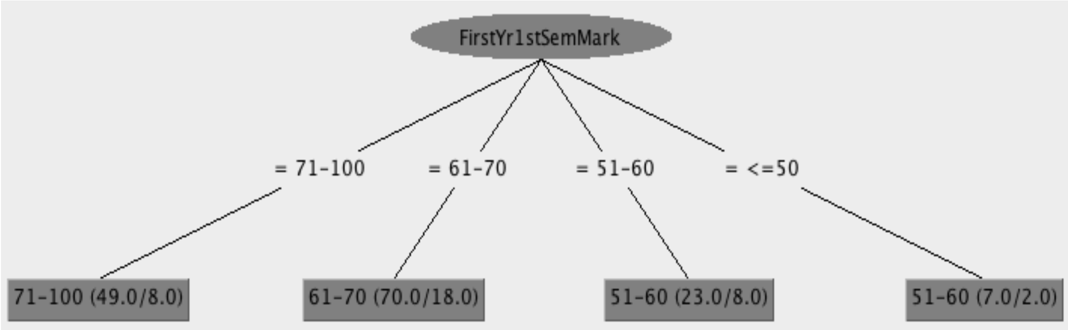


Figure 8.3 J48 rule for model 3 (institutional internal database based model with students' first semester marks).

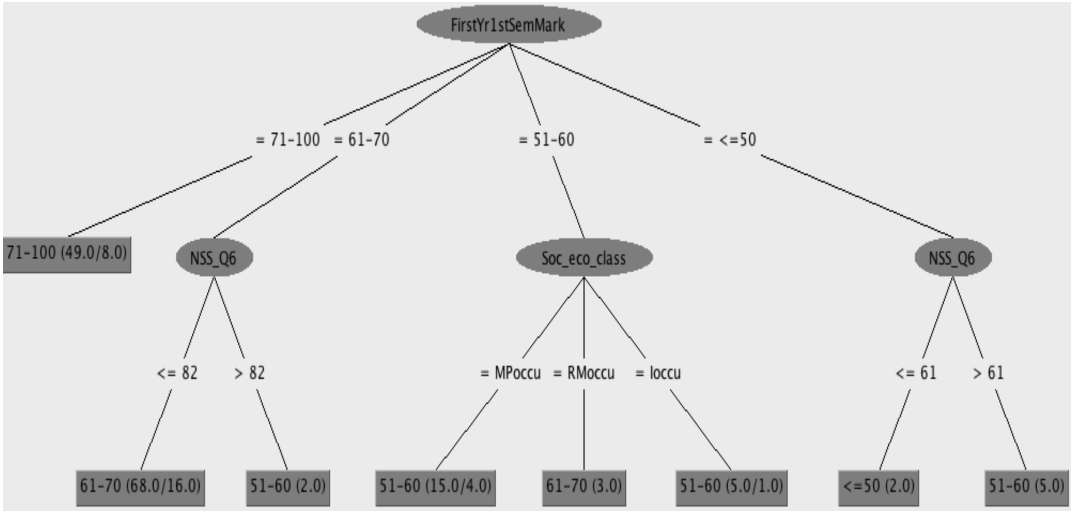


Figure 8.4 J48 rule for model 4 (institutional internal database and external open dataset based model with students' first semester marks).

The summaries of the classification models are presented in Table 8.6 (for model 1 and model 2) and Table 8.7 (for model 3 and model 4) with model accuracy. Additionally, both of the tables present class wise True Positive (TP[76]) rate, False Positive (FP[77]) rate, precision[78] and recall[79] value for each model.

[76] TP refers to the number of instances correctly classified as belonging to the positive class.

[77] FP refers to the number of instances incorrectly classified to the positive class.

[78] The Precision means the proportion of the instances which are truly in class x among all those which were classified as class x.

[79] Recall is defined as the number of true positives divided by the total number of elements that actually belong to the positive class.

Table 8.6 Summary of the classification model 1 (institutional internal database based model without students' first semester marks) and model 2 (institutional internal database and external open dataset based model without students' first semester marks).

| Model Name | Class | TP Rate | FP Rate | Precision | Recall | F-Measure | Model accuracy |
|---|---|---|---|---|---|---|---|
| Model 1 | 71-100 | 0.417 | 0.188 | 0.513 | 0.417 | 0.46 | 46.98% |
| | 61-70 | 0.769 | 0.69 | 0.463 | 0.769 | 0.578 | |
| | 51-60 | 0 | 0.017 | 0 | 0 | 0 | |
| | 41-50 | 0 | 0 | 0 | 0 | 0 | |
| **Weighted Average** | | **0.47** | **0.365** | **0.367** | **0.47** | **0.40** | |
| Model 2 | 71-100 | 0.563 | 0.238 | 0.529 | 0.563 | 0.545 | 52.35% |
| | 61-70 | 0.662 | 0.393 | 0.566 | 0.662 | 0.61 | |
| | 51-60 | 0.276 | 0.092 | 0.421 | 0.276 | 0.333 | |
| | 41-50 | 0 | 0.021 | 0 | 0 | 0 | |
| **Weighted Average** | | **0.523** | **0.267** | **0.499** | **0.523** | **0.507** | |

Table 8.7 Summary of the classification model 3 (institutional internal database based model with students' first semester marks) and model 4 (institutional internal database and external open dataset based model with students' first semester marks).

| Model Name | Class | TP Rate | FP Rate | Precision | Recall | F-Measure | Model accuracy |
|---|---|---|---|---|---|---|---|
| Model 3 | 71-100 | 0.854 | 0.079 | 0.837 | 0.854 | 0.845 | 74.50% |
| | 61-70 | 0.8 | 0.214 | 0.743 | 0.8 | 0.77 | |
| | 51-60 | 0.621 | 0.083 | 0.643 | 0.621 | 0.632 | |
| | 41-50 | 0 | 0.014 | 0 | 0 | 0 | |
| **Weighted Average** | | **0.745** | **0.136** | **0.719** | **0.745** | **0.731** | |
| Model 4 | 71-100 | 0.854 | 0.079 | 0.837 | 0.854 | 0.845 | 76.51% |
| | 61-70 | 0.815 | 0.214 | 0.746 | 0.815 | 0.779 | |
| | 51-60 | .69 | 0.075 | 0.69 | 0.69 | 0.69 | |
| | 41-50 | 0 | 0 | 0 | 0 | 0 | |
| **Weighted Average** | | **0.765** | **0.134** | **0.729** | **0.765** | **0.747** | |

The result presented in Table 8.6 shows that the model solely based upon institutional internal databases (model 1) achieved 46.98%, 37%, 47% and 40% accuracy, precision, recall and F-measure, respectively. On the other hand, the model based on institutional internal databases and external open data sources (model 2) attained a model accuracy of 52.35%, precision of 50%, recall of 52% and F-measure of 50%. From the model output, it can be noted that inclusion of external data in the predictive model enriched the prediction accuracy by 5.37% compared to model 1 (model solely based upon institutional internal databases). In addition, the other three measure (precision, recall and F-Measure) values are also higher in model 2 compared to model 1. Collectively, it suggests that model 2 performs better than model 1. Furthermore, the result of the t-test also suggests that model 2 is significantly better in predicting students' first year final marks than model 1 ($t = -11.03$ and $p<0.05$). Therefore, it can be said that information in external open data sources has the ability to improve the performance for predicting students' marks in their first year of study.

Furthermore, when we included students' current academic performance (first semester marks) in both predictive models (model 1 and model 2), the predictive models performance improved. An accuracy of 74.50% was achieved for the model solely based upon institutional internal databases (model 3) and 76.51% for the model using institutional internal databases and commonly available external open data sources (model 4). In addition, both of the models attained higher precision, recall and F-Measure values. Model 3 reached 72% precision, 75% recall and 73% F-measure, whilst, model 4 achieved 73% precision, 77% recall and 75% F-measure of the model performance. Therefore, from the results, it can be claimed that model 4 performs better than model 3, as model 4 achieved higher precision, recall and F-Measure values compared to that of model 3. Furthermore, the t-test results confirmed the performance difference between model 3 and model 4 are significant, which refers that model 4 performs significantly better than model 3 in predicting students' first year final marks ($t=-8.81$ and $p<0.05$).

It is perceived that after adding students' first semester marks in both prediction models (model 1 and model 2), the prediction performance increased remarkably. For model 1, the overall prediction accuracy increased from 46.98% to 74.50%, and for model 2, it increased from 52.35% to 76.51%.

In addition, both of the models (model 3 and model 4) achieved higher precision and recall values compared to model 1 and model 2. Therefore, it can be strongly said that students' first year final mark is highly correlated with students' current academic performance (first semester mark). However, this is not quiet surprising result because students' first year final marks are calculated by incorporating a certain percentages (around 50%) of students' first semester marks. Therefore, it can be speculated that students' first semester marks has influenced in predicting students' first year final marks. On the other hand, according to the recommendation of Sehgal *et. al.* (2012), the best performing model is the model that can provide higher accuracy, recall and precision values. Hence, we can say that model 4 is the best model among all the four models as it has provided the highest accuracy, recall and precision values.

The results of the current study support Kember (1995) and Kovacic (2010), where the authors stated that background characteristics are not good predictors of final outcomes. They reported that by using only background variables, the predictive models do not perform very well in predicting students' performance. Kovacic reported including only background variables in the classification model, prediction accuracy could be obtained around 60% or less. Also, the results support Kotsiantis *et al.* (2004), Cortez and Silva (2008) and Kovacic and Green (2010), where they stated that students' previous academic performance is highly correlated with their final marks.

The current study provides evidence that institutional external open data sources can be used in predicting students' academic performance. However, the current study has some limitations, which may affect the study output. This study used an unbalanced dataset of students for each mark group (the 61%-70% mark group consisted of the highest number of trials and the 41%-50% mark group consisted of the lowest number of trials compared to the other mark groups). We assume that all the classification models did not perform well in classifying students for the 41%-50% mark groups due to the lowest number of trials in this group. Therefore, it is recommended to include enough number of trials (preferably balance) for each class in future studies. In addition, this study used students' self reported data, which may lead to some wrong information and may influence model output. Similar to the previous experiment (presented in Chapter 7), in future the higher education institutions would allow access to detailed information of their students. This may reduce the dependency to use self-reported questionnaires and allow the development of a more robust model.

## 8.4 Summary

This chapter presents model development approaches that can be used in practical settings to predict students' academic performance using institutional internal and available external open data sources. Moreover, this chapter presents experiment 2, which tests hypothesis 4 regarding the sufficiency of existing external open data sources to predict students' first year mark. This experiment applied a decision tree algorithm for the prediction model development and used linked data technologies to integrate data from disparate data sources (institutional internal and external data sources). The result of the experiment shows that models based on institutional internal databases and external open data sources perform better than models based on only institutional internal databases. Also, the result strongly supports that students' current academic performance (first semester mark) is the best predictor in predicting students' mark. Among other predictors, A level points, and NSS results are also highly recommended for predicting students' marks in their first year of study.

Therefore, the experiment result underlines the importance of linked open data sources in developing such types of predictive models. We envisage that results such as the ones described in this experiment may increasingly improve the design of future students' predictive models to predict students' academic performance.

# Chapter 9: Discussion

This chapter illustrates the evidence for the hypotheses presented in this thesis. First, it provides an overview of methodologies of the two experiments and then, discusses the findings based on each hypothesis. This discussion suggests that there is sufficient evidence to support all four hypotheses. At the end, it states the limitations of the study.

## 9.1 Overview of the Methodology

In earlier two chapters, two experiments with their methodologies and results are presented. These experiments were conducted to test the hypotheses stated in Chapter 1. The first experiment explored a new approach to develop student predictive models to identify at-risk students that rely on data commonly available in institutional internal databases and external open data sources instead of questionnaires used in the traditional student predictive models. The second experiment was to develop student predictive model to predict students' academic performance in their first year of study using institutional internal databases and external open data sources.

In the first experiment, three predictive models were developed. The first model is a survey-based model and it includes all the independent variables (institutional internal database items and questionnaire items) considered in a previous study by Pascarella *et al.* (1983) to develop the student predictive models to find out students who are most likely to be at-risk in their first year of study. Their survey-based study is derives from Tinto's student retention theory (1975). Pascarella and Terenzini's Institutional Integration Scale (1980) was used to measure various dimensions identified by Tinto (1975, 1993) as corresponding to the likelihood of dropout in the first year of study. The second model includes only the subset of variables from the first model (survey-based model), which are commonly available in the institutional internal databases to perceive how the model performs with only the available data in HEI. Finally, the third model includes all the variables from the second model (only institutional internal database variables) and includes new variables from external open data source as the replacement of the traditional questionnaire items/variables from the first model. Data was analysed using categorical principle component analysis and binary logistic regression. To validate the models, repeated hold out method was applied where, 70% of the data was used for training the model and 30% of the data was used for testing the model. These three models were compared with respect to their overall prediction accuracy, sensitivity, specificity, type I error rate and type II error rate to identify the best performing model among them.

The purpose of the second experiment was to explore a new way of developing student predictive model to predict students' performance in their first year of study using institutional internal databases and external open data sources. In the first stage of the experiment, two predictive models were developed. The first model includes all the independent variables available in the institutional internal databases based on a previous study by Yadev and Pal (2012) to predict students' academic performance. The second model includes all the variables from the first model and includes new variables from external open data sources. Two models were compared based on their total model accuracy, precision, recall and F-Measure values. Later, we extended these two predictive models by adding students' first semester marks in order to observe the effects of adding students' first semester marks in the model performance. Moreover, this allowed us to analyse the effect of external open data sources on the both predictive models before and after adding current academic performance (first semester mark) in the predictive models. These two models were also compared based on their total model accuracy, precision, recall and F-Measure values. In selecting significant attribute, Ahmad and Dey's (2004) proposed significant attribute evaluator method was used. J48 decision tree was used to develop the predictive models and 10-fold cross validation method was used to validate the models.

In order to collect all the student specific data that are available in the institutional internal databases as well as those are collected through questionnaire in the survey-based model, an online questionnaire session was conducted over a 12-week period at the University of Southampton (UK) as we did not have university database access permission at that time. In total, 149 students participated in the questionnaire session. A linked data experimental platform has been developed which is able to convert raw data to RDF and also able to integrate data from various data sources to make a final set of data for further analysis.

## 9.2 Discussion

### 9.2.1 Hypothesis 1

***Hypothesis 1*** claims that it is possible to provide accurate/improved student prediction models by combining internal databases and external open data sources. The results of the first experiment provide evidence to support hypothesis 1 as the student predictive model to find out at-risk students in their first year of study using institutional internal databases and external open data sources performed the best among all the three predictive models. It achieved the highest model accuracy of 89.20%, sensitivity of 89.33% and specificity of 89.63%. The second best predictive model was the survey-based one with an overall model accuracy of 88.86%, sensitivity of 84.33% and specificity of 89.57%, and the lowest performing one was the solely institutional internal database based model, which attained 84.94% of the model accuracy, 61.33% of the sensitivity and 87.46% of the specificity. Though the specificity between the survey-based model (89.57%) and institutional internal databases and external open dataset based model (89.63%) are quite comparable. Nevertheless, with respect to the sensitivity, the latter model attained the highest sensitivity (89.33%), which is about five percent higher than the survey-based model (84.33%) and 28 percent higher than the model solely based upon institutional internal databases (61.33%). In addition, type I error (false positive[1]) and type II error rate (false negative[2]) are comparatively smaller for the model based on institutional internal databases and external open data sources (type I error, 10.37% and type II error, 10.67%) than the survey-based model (type I error, 10.43% and type II error, 15.67%)

[1] A false positive or type I error occurs when a student is mistakenly marked as "at-risk" instead of "not at-risk".

[2] A false negative or type II error occurs when a student is marked as "not at-risk" but in reality the student is "at-risk".

and solely institutional internal databases based model (type I error, 12.54% and type II error, 38.67%).

The consequence of type I error is that HEI need to arrange additional learning support for the misclassified students even though they do not need it. On the other hand, the consequence of type II error is that the misclassified students would not receive additional learning support provided to the students at-risk. Therefore, it is desirable to have both error rates as minimum as possible in the best performing predictive model. In this context, predictive model 3 seems satisfies this aim and performed better than other two predictive models.

The predictive model based on institutional internal databases and external open data sources (model 3) achieved quite high level of predictive validity when compared to previous efforts to model persistence. The base model by Pascarella *et al.* (1983) achieved 82.1% of the model accuracy using IIS questionnaires and institutional databases. A previous study of student retention conducted by Herzog (2005) achieved 77.4% of the model accuracy by applying logistic regression on institutional databases and National Student Clearinghouse dataset. In another study, Herzog (2006) achieved about 75% of the model accuracy using logistic regression, decision trees and neural network. In his study of student retention, Sujitparapitaya (2006) achieved about 80% of the model accuracy using logistic regression method. A subsequent retention study conducted by Delen (2010) using institutional internal databases achieved 74.26%-81.18% of the model accuracy by applying support vector machine, neural network, decision tree and logistic regression. Comparatively, we achieved an overall model accuracy of 89.20% using institutional internal databases and external open data sources by applying logistic regression. The accuracy of the model exceeds and compares favorably to findings of the previous studies. In addition, the finding of this research is consistent with past research regarding the importance of parents HE qualifications, students' A level points and their first semester marks in predicting vulnerable students (Tinto, 1993; Elkins *et al.* (2000); Ishitani, 2006; Hosch, 2008).

In addition, model 3 explained higher percentage of variance (59%) in the data compared to model 1 (54%) and model 2 (50%). The proportions achieved in the current study are quite comparable to some of the previous studies. For example, Allen and Nelson (1989) achieved from 44% to 53%, Milem and Berger (1997) achieved 41% , Berger (1997) achieved 42%, Berger and Braxton (1998) achieved 44% of the variance in retention. However, most of the studies based on Tinto's theory explained very low proportion of the variance in the retention. For example, Pascarella *et al.* (1983) attained 28.1%,  Fox (1986) attained 31%, Berger and Milem (1999a) attained 25%, Thomas (2000) attained 26% of the variance in the retention.

### 9.2.2 Hypothesis 2

***Hypothesis 2*** claims that linked data is efficient to support in building student prediction model when combining internal and external data sources. As institutional internal data and external open data are spread out in different sources in different format, it is challenging to integrate data from these separate data sources to make a single dataset to analyse. Linked data is found to be well suited in combining data from disparate data sources as it provides more expressivity of the data. In linked data, taxonomies, vocabularies and ontologies provide domain specific terms to describe classes in RDF and how they relate to one other. Moreover, it creates typed links between data from different data sources. Furthermore, linked data provides provenance and validation of data while integrating data from disparate data sources.

Although, linked data is efficient in integrating and interoperability of data from different data sources, we still are not getting the full benefit from it. We identified three main issues which hindrance from utilizing full advantage of linked data:

**Lack of data standardization:** It is noticed that the major issue in integrating data from multiple data sources is the lack of standardization in the data as most of the interested data are in 2 star (.xls) or 3 star (.csv) format[3]. This can be improved if data providers would publish their data in linked data format using standardized vocabularies and ontologies and allow their data available via a SPARQL endpoint.

**Real time data integration:** RDF data integration is done by loading all data into a single repository and querying the merged data locally. This is not feasible for technical reasons. The possible technical reason is that local copies

[3] http://www.w3.org/DesignIssues/LinkedData.html

are not up-to-date. In other words, to maintain up-to-date, data would require the local copy to be updated locally constantly even when changes occur in the main dataset, which is not feasible. This problem could be resolve if data provider make their data available via a SPARQL endpoint, which would permit end user to access up-to-date data within compromising the main dataset.

**Premature SPARQL:** In the context of statistical methods, SPARQL is still at the early stage. Different frameworks and tools, which are using SPARQL, have already implemented aggregate functions like MAX, MIN, AVG or SUM. Some of these extensions are found recognition in SPARQL 1.1[4]. An overview of proposed and implemented extensions can be found at the corresponding page in the W3C-Wiki[5]. More functions and extensions are required to be included in the SPARQL to provide flexibility to the application developments. For example, inclusion of more complex statistical methods, such as, principle component analysis, decision trees, different feature selection methods, logistic regression could improve the flexibility of the platforms of the current research study. However, inclusion of such methods in SPARQL would be highly complex and expensive. To avoid the complexity of the calculations Zapilko and Mathiak (2011) recommended to use the available tools to perform the analysis. Kiefer *et al.* (2008) extended SPARQL to SPARQL-ML (SPARQL Machine Learning) to perform different data mining techniques such as, classifications and clustering on linked data. They implemented SPARQL-ML as an extension to ARQ (the SPARQL query engine for Jena[6]). The authors believe that SPARQL-ML could serve as a standardized approach for data mining tasks on linked data. Therefore, it could be beneficial to make an extension of SPARQL, which includes all data mining methods available for all platforms to serve the purpose.

### 9.2.3 Hypothesis 3

***Hypothesis 3*** claims that Internal/external data sources can be used to compensate the lack of questionnaire data in building student prediction model. National Student Survey (NSS) result published in Unistats website was used to replace Institutional Integration Scale (IIS) questionnaires in experiment

[4]http://www.w3.org/TR/sparql11-query

[5] http://esw.w3.org/SPARQL/Extensions

[6] http://jena.sourceforge.net/

1. As IIS was used to measure different dimensions of student satisfaction and integration, we include a total of 16 questionnaire items from NSS which related to student faculty interaction, faculty concern for student development, student development and about their course among the 22 NSS common questionnaires for all subjects as a replacement of the IIS questionnaire. The results of the first experiment provide evidence that external data sources can be used instead of traditional questionnaire to build student predictive model as the model using institutional internal databases and external open data sources performs best compared to the survey-based model and solely institutional internal database based model. The results of the second experiment also support this hypothesis as including external open data (16 NSS questionnaire) in developing the student predictive model increases overall model performance. Moreover, parents' annual income, students' socio-economic status, participation neighborhood have been traditionally used in many predictive models to predict students' academic performance and typically this information are collected through student questionnaire. In this experiment, Office for National Statistics (ONS) published dataset (gross annual pay based on SOC2010) has been used to derive parents' annual income and socio economic class (based on SOC2010) of the student. In addition, participations neighborhood group has been derived using Higher Education Funding Council for England (HEFCE) published dataset to include in the model.

We counsel to take precaution in linking internal databases' item and external open dataset's item, as most of them are defined in different names or ids in the datasets. Particularly, we experienced difficulties while linking students' parental occupation provided by students (through online questionnaire) with the ONS published occupation list (SOC2010) to derive parents' annual income and students' socio economic group as they were defined in different names in their own datasets. For example, some students defined their parental occupation as "specialist nurse" but in ONS published dataset there is no occupation defined as "specialist nurse". In ONS occupation list, there is an occupation defined as "nurse". Therefore, we linked these two occupations assuming that they are same. Also, few students defined parental occupation as Higher Level Teaching Assistant (HLTA) while ONS defined as "special needs education teaching professionals". However, this has been a well-documented issue and known as terminological or semantic heterogeneity or co-reference problem in the area of ontology mapping (Cui *et al.*, 2001; Mitra and Wiederhold, 2002). Several efforts have been taken to resolve the heterogeneity problem but still remain inadequate to automatically resolve the problem, often requiring manual intervention (Kalfoglou and Schorlemmer, 2003). Therefore, attention needs to be taken during the linking of the disparate datasets.

### 9.2.4 Hypothesis 4

***Hypothesis 4*** claims that it is possible to predict students' mark using institutional internal databases and external open data sources. The results of the second experiment provide evidence that it is possible to predict students' mark using institutional internal databases and external open data sources. The results also suggest that the predictive models using institutional internal databases and external open data sources perform better compared to the models solely based upon institutional internal databases. Additionally, it is noteworthy that while selecting significant variables for the model development, NSS five questionnaires (Q2, Q6, Q9, Q5 and Q8) positioned top second, third, fourth, fifth and sixth with highest score value. From the experiment results, it is found that adding students' first semester marks in both predictive models (model based on solely institutional internal datasets, and model using institutional internal databases and commonly available external open data sources) increased the prediction performance remarkably. The prediction performance for solely based upon institutional internal databases increased from 46.98% to 74.50%, and 52.35% to 76.51% for institutional internal databases and external open data sources based model, which positioned current academic performance (first semester mark) as the most important variable in predicting students' first year's final mark. This finding is consistent with past research regarding the importance of current academic performance (first semester mark) in predicting students' final marks (Kam and Ch'ng, 2009). In addition, A level points (previous academic performance) also contributed the most in predicting students' final marks among other variables. This finding also supports several previous studies conducted in predicting students' performance such as, Al-Radaideh *et al.* (2006), Kotsiantis *et al.* (2004) and Kovacic and Green (2010).

It worth highlighting that while included data from external open data sources in developing predictive models, the overall model accuracies increased in both cases (with and without first semester marks). The model accuracy of the predictive model without first semester marks increased by 5.37% while included data from external open data sources. Also, it is well mentioned that including first semester marks in developing predictive models, the overall model accuracy increased in both predictive models but still remain 2.01% higher in the institutional internal databases and external open data sources based model than the model solely based upon institutional internal databases.

The predictive models based on institutional internal databases and external open data sources (model 2 and model 4) have achieved a good level of predictive performance when compared to previous efforts to predict academic performance. In a previous study conducted by Kotsiantis *et al.* (2004), used key demographic variables and assignment marks to predict student's performance and achieved from 58.84% to 64.47% of the model accuracy. Vandamme *et al.* (2007) reported from 40.63% to 57.35% of an overall model accuracies in their study of students' performance prediction. Kovacic (2010) used students' enrolment data and achieved from 59.4% to 60.5% of the model accuracy in predicting students' performance. Yadav and Pal (2012) obtained data from institutional databases to predict students' academic performance and they obtained from 62.22% to 67.77% of the model accuracy. In another study, Yadev *et al.* (2011) attained from 45.83% to 56.25% of the prediction accuracy. Comparatively, we achieved 52.35% (without first semester marks) and 76.51% (with first semester marks) of the model accuracy by including external data sources in the development of the predictive models. This results compares favourably to findings of the previous studies.

## 9.3 Limitations of the Present Research Study

According to this research study, there are several limitations that need to be addressed. Each of the limitations is attributed as follows:

- The sample size of the current research study is 149. Though, hypotheses of this research were tested adequately with the current sample size, it would be better to have a bigger sample size to derive more robust findings using the proposed approach in this research.

- Both of the experiments in this research contained an unbalanced dataset. In experiment 1, 15% of the students were in the at-risk student group and 85% of the students were in the not at-risk student group. In experiment 2, 32.89% of the students were in the 71%-100% marks group, 43.62% of the students were in the 61%-70% marks group, 19.46% of the students were in the 51%-60% marks group and 4.02% of the students were in the 41%-50% marks group. It was difficult to have a balanced dataset as we did not have access permission to the University of Southampton's students' databases. Therefore, we had to rely on students who willingly participate in the questionnaire session. Though, the current research study overcome this issue by considering a number of evaluation criteria, such as, accuracy, specificity, sensitivity, type I error rate and type II error rate to validate the model performance in experiment 1. On the other hand, accuracy, recall, precision and F-measure values were considered to validate the model performance in experiment 2. It is recommended to use balanced dataset in future research studies to develop more robust predictive models.
- This research focused only on student progression during the freshman year, and therefore, student progression in subsequent years was not assessed. In future investigations that looking into the subsequent years could prove more informed findings in this area.
- We could not classify the real dropout students, because, we did not have means to contact them and we did not have access permission to the University of Southampton's student database for this purpose. However, investigation including the data of the real dropout students could able to develop robust predictive model.

- For the purpose of this research, all the students' specific data provided by students themselves during the online questionnaire session, which may lead some wrong information. Students self reported data may influence model output. However, integration of authentic data can elevate this issue.
- The current research study is conducted at a single institution, University of Southampton. Therefore, the findings of this research may not be generalizable to other institutions. Future research on integration of multiple institutional datasets is required to develop robust and generalized predictive models that could be effectively useful in the higher education institutions.
- Finally and most importantly, due to the current limitation of SPARQL to manage complex statistical methods, data analysis part of this research was completed using SPSS and WEKA software. Extension of SPARQL including of more complex statistical methods in SPARQL, such as, principle component analysis, decision trees, different feature selection methods, logistic regression could improve the flexibility of the platforms.

# Chapter 10: Conclusions

The final chapter of this thesis summarizes the purpose of this PhD research in section 10.1. Section 10.2 outlines the major findings of this research and section 10.3 makes recommendations for future research based on the findings of the current research. Finally, section 10.4 concludes this thesis chapter with some final thoughts.

## 10.1 Purpose of this PhD Research

Student retention and progression remains a central policy issue that demands active consideration by policymakers and other higher education stakeholders across the globe. There are many student retention models developed by many researchers. Over the course of the last thirty years, many student retention models have been proposed. For instance, Tinto (1975; 1993) has developed a longitudinal internationalist model of student retention, which has been widely accepted by the research community as a good working theory. Tinto's theory is dynamic and views student retention decisions largely as the results of interactions between the student, the academic and social systems of the institution (Tinto, 1975; Tinto, 1987; 1993). Institutional Integration Scale (IIS), which is a set of questionnaire developed by Pascarella and Terenzini (1980), has been popularized to measure various dimensions of Tinto's model. Most of the studies use Institutional Integration Scale (IIS) to develop and/or test Tinto's retention model. Therefore, research in student retention and progression to completion is typically conducted through survey, where researchers collect data through questionnaires and interviewing students. The major issues of survey-based study are the potential for low response rates, high cost and administrative cumbersome.

Nowadays, the volume of data collected by higher education institutions has increased more than ever before. Alongside this increased volume of data, repositories are helping in efficiently storing and accessing this data. Apart from the institutional data there is a large amount of external open data available on the web. Furthermore analytics are providing increased opportunities to higher education sector in making informed decision. At the same time, linked data technologies provide efficient data integration from different sources.

The purpose of this PhD research is to examine the capability of linked data technologies and the sufficiency of existing open data sources in supporting student retention, progression and completion. A further aim is to explore a new approach to build student predictive models through integrating linked data sources that are internal or external to higher education institutions, without having to rely on questionnaire data that have been essential in existing models. In addition, this PhD research study explore a new approach to develop a student performance prediction model using institutional internal and external linked open data sources to predict students' first year performance, to arrange support for the poor performing students in their study to progress and complete their study successfully.

## 10.2 Major Findings

Based on the results of the experiments, we can specify the major findings of this research study as follows:

- Student predictive model based on institutional internal databases and external open data sources (model accuracy 89.20%, sensitivity 89.33%, specificity 89.63%, type I error 10.37%, type II error 10.67% and variation explained 59%) could perform as well as or even out-perform the traditional survey-based predictive model (model accuracy 88.86%, sensitivity 84.33%, specificity 89.57%, type I error 10.43%, type II error 10.67% and variation explained 54%) and the model solely based upon institutional internal databases (model accuracy 84.94%, sensitivity 61.33%, specificity 87.46%, type I error 12.54%, type II error 38.67% and variation explained 50%). Therefore, it can be said that it is possible to provide accurate or improved predictive model in combining institutional internal databases and external open data sources.

- The available internal and external open data sources can be used to compensate the lack of questionnaire data in building student prediction model, such as, Unistats's published National Student Survey data, Office for National Statistics' published annual gross income data, Higher Education Funding Council for England's published participation neighbourhood data.

- Linked data is efficient to support in building student prediction model when combining internal and external open data sources. Linked data is efficient in terms of data integration as it provides more expressivity of data. In addition, linked data reduces cost by saving time and money require for organising survey and manual integration of data from different sources. Though, there are some issues need to be addressed to get full benefit from linked data technologies, such as, the lack of standardization in the current data and SPARQL's ability to perform complex statistical methods.

- From the result of the second experiment, it is found that predictive models using institutional internal databases and external open data sources perform best with the model accuracies 52.35% (without first semester mark) and 76.51% (with first semester mark) compared to the models solely based upon institutional internal databases 46.98% (without first semester mark) and 74.50% (with first semester mark) respectively in predicting students' mark in their first year of study. Therefore, it can be said that combining institutional internal databases and external open data sources can better support in predicting student's mark.

## 10.3 Recommendation for Future Research

Based on the literature reviewed and the two large empirical studies presented and discussed in this thesis, the following recommendations are made for the future research:

- The current investigation is based on only one higher education institution (University of Southampton) in the UK and the outcome of this investigation is found significant. Future research can be conducted at other higher education institutions in the UK or outside the UK to confirm the findings of this investigation.

- The sample size of the current study is 149. Though, the hypotheses of the current research tested adequately with the current sample size, a larger sample of first-year students is recommended to make the predictive models more robust.

- The current research considered student progression during the first year. Since, the pattern of influences may not be the same for other students in other academic years, future research should focus upon student progression in subsequent years. Houston *et al.*, (2003) found that while non-progression was greatest in the first year, it was still an issue in subsequent years.

- The current research achieved about 50% to 59% of the variance in the dependant variable. This indicates that at least some important predictors of student retention may not be properly specified by the theory. Thus, more research is needed to identify these predictors. As noted earlier, this would require a larger sample and preferably more than one institution. In addition, information available from VLE can be included in retention study.

- The current research focused only on factors drawn from Tinto's theory. Future research should investigate additional factors from external open data sources.

- Experiment 1 applied logistic regression approach and repeated hold-out method to develop and validate the student predictive models. It did not make comparisons with any other approach, for example: Bayesian classification, support vector machine, neural network, decision trees. Such comparison may improve the model robustness. Therefore, it is recommended to apply such approaches in future study.

- Similarly, Experiment 2 applied decision tree approach and 10-fold cross validation method to develop and validate the student predictive models. It did not make comparisons with any other approach, for example: Bayesian classification, support vector machine, neural network, logistic regression. Such comparison may improve the model robustness. Therefore, it is recommended to apply such approaches in future study.
- Future research can consider investigating the capabilities of linked data technologies and external open data sources to support other higher education challenges.
- The SRPC ontology which is developed in this thesis is in its initial stage, the future work consider to develop a complete SRPC ontology and make it available online.

## 10.4 Concluding Remarks

Student retention and progression remains one of the top issues to be considered by policy makers and those engaged in higher education worldwide. Retention literature is replete with evidence that traditionally retention research is survey-based where researchers use questionnaire to collect students' data multiple times during their study life to develop student predictive models to find out at-risk students in their programmes of study. The major issues with survey-based study are the potentially low response rates and costly. Moreover, organizing questionnaires in multiple times can be overbearing for higher education institutions.

This research sought to examine the capability of linked data technologies and the sufficiency of existing external open data sources in supporting student retention, progression and completion. Moreover, the aim has been to explore a new method of developing student predictive models through integrating linked open data sources that are internal or external to higher education institutions, without having to rely on questionnaire data that have been essential in existing models. The results of the experiments confirm that prediction models using institutional internal databases and external open data sources perform better in predicting at-risk students in their first year of study as well as in predicting students' first year final mark than its counterpart models. The finding has been of special interest for the researcher who has retention responsibilities at the university and can see the potential of further utilizing the existing linked open data sources with a larger sample of first-year students to seek to replicate the results. If this finding is further validated, institutions would do well to use it to add to their collection of predictive tools. We believe that results and approach such as the ones described in this thesis may increasingly improve the design of future students' predictive models to support students to perform better in their academic trajectory.

# References

1st International Conference on Learning Analytics and Knowledge 2011. Banff, Alberta, https://tekri.athabascau.ca/analytics/.

Abdi, H. 2003. Factor Rotations in Factor Analyses. In M. Lewis-Beck, A. Bryman, and T., *Futing, editors, Encyclopedia of Social Sciences Research Methods. Sage, Thousand Oaks, CA.*

American Council on Education (ACE), American Association of State Colleges and Universities (AASCU), American Association of Community Colleges (AACC), Association of American Universities (AAU), National Association of Independent Colleges and Universities (NAICU), National Association of State Universities & Land Grant Colleges (NASULGC) 2006. A Letter to Our Members: Next Steps "Addressing the Challenges Facing American Undergraduate Education", http://www.harford.edu/irc/facultyresources/Big6_ResponseLetter_Final.pdf.

Ahmad, A. and Dey, L. 2004. A feature selection technique for classificatory analysis. *Pattern Recognition Letters*, 26(1), 43-56, http://sci2s.ugr.es/keel/pdf/specific/articulo/Ahmad_A_Feature_2004.pdf.

Al-Radaideh, Q. A., Al-Shawakfa, E. M. and 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).*

Allemang, D. and Hendler, J. 2008. *Semantic Web for the Working Ontologist: Eective Modeling in RDFS and OWL*, Morgan Kaufmann Publishers Inc., San Francisco, USA.

Allen, D. F. and Nelson, J. M. 1989. Tinto's model of college withdrawal applied to women in two institutions, *Journal of Research and Development in Education,* 22(3), 1-11.

Anderson, T., Howe, C., Soden, R., Halliday, J. and Low, J. 2001. Peer interaction and the learning of critical thinking skills in further education students, *Journal of Instructional Science,* 29 (1), 1-32, http://www.springerlink.com/content/n37m8381274311578/.

Arnold, K. E. 2010. Signals: Applying Academic Analytics, EDUCAUSE Quarterly (EQ) Magazine, 33(1), http://www.educause.edu/ero/article/signals-applying-academic-analytics.

Astin, A. W. 1975. *Preventing students from dropping out*, San Francisco: Jossey-Bass.

Astin, A. W. 1984. Student involvement: A developmental theory for higher education, *Journal of College Student Personnel,* 25(4), 297–308.

Astin, A. W. 1993. *What matters in college: Four critical years revisited*, San Francisco:Jossey-Bass.

Bailey, B. L., Bauman, C. a. and Lata, K. A. 1998. Student retention and satisfaction: The evolution of a predictive model. Minneapolis, MN: Association for Institutional Research. (ERIC Document Reproduction Service No. ED 424 797).

Bailey, M. and Borooah, V. K. 2007. Staying the Course: An Econometric Analysis of the Characteristics most Associated with Student Attrition Beyond the First Year of Higher Education, Department for Employment and Learning Research Report, http://www.delni.gov.uk/staying_the_course.pdf

Bean, J. P. and Metzner, B. S. 1985. A conceptual model of non-traditional undergraduate student attrition, *Review of Educational Research,* 55(4), 485-540.

Bean, J. P. 1980. Dropouts and turnover: The synthesis and test of a causal model of student attrition, *Research in Higher Education,* 12(2), 155-187.

Bean, J. P. 1985. Interaction effects based on class level in an explanatory model of college student dropout syndrome, *American Educational Research Journal,* 22(1), 35-65.

Bean, J. P. and Eaton, S. B. 2000. A psychological model of college student retention. In J. M. Braxton (Ed.), *Reworking the departure puzzle: New theory and research on college student retention*. Nashville: University of Vanberbilt Press.

Bean, J. P. and Eaton, S. B. 2002. The psychology underlying successful retention practices, *Journal of College Student Retention,* 3(1), 73-89, http://homepages.se.edu/native-american-center/files/2012/2004/The-Psychology-Underlying-Successful-Retention-Practices.pdf.

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F. and Stein, L. A. 2004. *OWL web ontology language reference, Recommendation, W3C,* http://www.w3.org/TR/owl-ref/.

Beer, C., Clark, K. and Jones, D. 2010. Indicators of engagement, curriculum, technology and transformation for an unknown future, *In the Proceedings of ASCILITE 2010,* 75-86, http://www.ascilite.org.au/conferences/sydney10/procs/Beer-full.pdf.

Beer, J. 2010. Key Issues facing UK HEIs: Staff quality, experience, skills and skills gaps, JISC report, http://www.jisc.ac.uk/media/documents/aboutus/strategy/janet beer essay.pdf.

Berger, J. B. and Milem, J. F. 1999a. The role of student involvement and perceptions of integration in a causal model of student persistence, *Research in Higher Education,* 40(6), 641-664.

Berger, J. B. and Milem, J. F. 1999b. The role of student involvement and perceptions of integration in a causal model of student persistence, *Research in Higher Education,* 40(6), 641-664.

Berger, J. B. 1997. Students' sense of community in residence halls, social integration, and first-year persistence, *Journal of College Student Development,* 38(5), 441-452.

Berger, J. B. and Braxton, J. M. 2000. Revising Tinto's Interactionalist theory of student departure through theory elaboration: Examining the role of organizational attributes in the persistence process, *Research in Higher Education,* 39(2).

Berger, J. B. and Braxton, J. M. 1998. Revising Tinto's internationalist theory of student departure through theory elaboration: Examining the role of organizational attributes in the persistence process, *Research in Higher Education,* 39(2), 103–120.

Berners-Lee, T. 2009. Linked Data - Design Issues, http://www.w3.org/DesignIssues/LinkedData.html.

Berners-Lee, T. 1998. RDF and Relational Databases. http://www.w3.org/DesignIssues/Overview.html.

Berners-Lee, T., Hendler, J. and Lissila, O. 2001. The semantic web: Scientific american, *Scientic American,* 284(5), 28-37.

Bharadwaj, B. K. and Pal, S. 2011a. Data Mining: A prediction for performance improvement using classification, *International journal of computer Science and Information security (IJCSIS),* 9(4), 136-140.

Bharadwaj, B. K. and Pal, S. 2011b. Mining Educational Data to Analyze Students' Performance, *International Journal of Advance Computer Science and Applications (IJACSA),* 2(6), 63-69.

Bichsel, J. 2012. Analytics in Higher Education: Benifits, Barriers, Progress, and Recommendations (Research Reort), Louisville, CO: EDUCAUSE Centre for Applied Research, http://www.educause.edu/ecar.

Biggs, J. and Tang, C. 2007. *Teaching for quality learning at university, 3rd edition,* Buckingham, UK: The Society for Research into Higher Education and Open University Press. http://docencia.etsit.urjc.es/moodle/pluginfile.php/18073/mod_resource/content/0/49657968-Teaching-for-Quality-Learning-at-University.pdf.

Department for Business, Innovation and Skills (BIS), 2009. Higher ambitions: The future of universities in a knowledge economy, http://dera.ioe.ac.uk/id/eprint/9465.

Bizer, C., Cyganiak, R. and Heath, T. 2007. How to Publish Linked Data on the Web, Tutorial. http://www4.wiwiss.fu-berlin.de/bizer/pub/linkeddatatutorial/.

Bizer, C., Heath, T. and Berners-Lee, T. 2009. Linked Data – The Story So Far, *International Journal on Semantic Web and Information Systems,* 5 (3), 1-22.

Borst, W. N. 1997. Construction of engineering ontologies for knowledge sharing and reuse, Center for Telematics and Information Technology, Enschede, Netherlands, http://doc.utwente.nl/17864/1/t0000004.pdf.

Boston, W., Diaz, S. R., Gibson, A. M., Ice, P., Richardson, J. and Swan, K. 2009. An Exploration of the Relationship Between Indicators of the Community of Inquiry Framework and Retention in Online Programs, *Journal of Asynchronous Learning Networks,* 13(3), https://www.ideals.illinois.edu/bitstream/handle/2142/18712/Boston%2520JALN13(3).pdf?sequence=2.

Braxton, J. M., Milem, J. F. and Sullivan, A. S. 2000. The influence of active learning on the college student departure process: Toward a revision of Tinto's theory, *Journal of Higher Education,* 71(5), 569-590.

Braxton, J. M. 2002. *Introduction: Reworking the student departure puzzle. In J. Braxton (Ed.), Reworking the student departure puzzle (2nd ed.),* Nashville: Vanderbilt University Press.

Braxton, J. M. 2000. *Reworking the student departure puzzle, First edition*, Vanderbilt University Press. http://books.google.co.uk/books?id=WF8itWof7aIC&printsec=frontcover&source=gbs_ge_summary_r&cad=0 - v=onepage&q&f=false.

Braxton, J. M., Duster, M. and Pascarella, E. T. 1988. Causal modeling and path analysis: An introduction and an illustration in student attiition research, *Journal of College Student Development,* 29, 263-272.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. 1984. *Classification and Regression Trees,* Wadsworth Int., CA.

Bridges, D. 2000. Back to the Future: the higher education curriculum in the 21st century, *Cambridge Journal of Education,* 30(1), 37-55, http://www.itslifejimbutnotasweknowit.org.uk/files/CPLHE/CJEBridgesCurric.pdf.

Britain, S. and Liber, O. 1999. A Framework for Pedagogical Evaluation of Virtual Learning Environments, http://www.jisc.ac.uk/uploaded_documents/jtap-041.doc.

Broekstra, J., Kampman, A. and van Harmelen, F. 2002. Sesame: A generic architecture for storing and querying RDF and RDF Schema, First International Semantic Web Conference (ISWC'02). Sardinia, Italy.

Browitt, A. and Walker, L. 2007. Retention and Widening Participation in the Faculties of Sciences and Engineering, University of Glasgow, http://www.suttontrust.com/research/retention-and-widening-participation-glasgow/.

Brunsden, V., Davies, M., Shevlin, M. and Bracken, M. 2000. Why do HE students drop out? A test of Tinto's model, *Journal of further and Higher Education,* 24(3), 301-310.

Brynjolfsson, E., Hitt, L. and Kim, H. 2011. Strength in Numbers: How Does Data--Driven Decisionmaking Affect Firm Performance?, Social Science Research Network, http://www.a51.nl/storage/pdf/SSRN_id1819486.pdf.

Cabrera, A. F., Castaneda, M. B., Nora, A. a. and Hengstler, D. 1992a. The convergence between two theories of college persistence, *Joumal of Higher Education,* 63(2), 143-164.

Cabrera, A. F., Nora, A. and Castaneda, M. B. 1992b. The role of finances in the persistence process: A stinctural model, *Research in Higher Education,* 33(5), 571-593.

Caison, A. L. 2007. Analysis of Institutionally Specific Retention Research: A Comparison Between Survey and Institutional Database Methods, *Research in Higher Education,* 48(4).

Campbell, J. P. and Oblinger, D. G. 2007. Academic analytics, EDUCAUSE Center for Applied Research, http://net.educause.edu/ir/library/pdf/pub6101.pdf.

Campbell, J. P., DeBlois, P. B. and Oblinger, D. G. 2007. Academic Analytics: A New Tool for a New Era, EDUCAUSE, https://net.educause.edu/ir/library/pdf/ERM0742.pdf.

Carroll, D. C. 1989. College persistence and degree attainment for 1980 high school graduated: Hazards for transfers, stopouts and part-timers, Washington, D.C: National Center for Education Statistics.

Carroll, J. and Appleton, J. 2001. Plagiarism: A good practice guide, JISC report, http://www.jisc.ac.uk/uploaded_documents/brookes.pdf.

Chen, P. D., Gonyea, R. and Kuh, G. 2008. Learning at a Distance: Engaged or Not?, *Innovate: Journal of Online Education,* 4(3).

Cooper, A. 2012. CETIS Analytics Series, Vol. 1, No. 5, What is Analytics? Definition and Essential Characteristics, http://publications.cetis.ac.uk/wp-content/uploads/2012/11/What-is-Analytics-Vol1-No-5.pdf.

Cortez, P., and 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.

Crosling, G., Heagney, M. and Thomas, L. 2009. Improving Student Retention in Higher Education: Improving Teaching and Learning, *Australian Universities*' *Review,* 51(2), 9-18, http://www.universityworldnews.com/filemgmt_data/files/AUR_51-02_Crosling.pdf.

Cui, Z., Jones, D. and O'Brien, P. 2001. Issues in Ontology-based Information Integration. *IJCAI – Seattle, USA.*

Davenport, T. H., Harris, J. G. and Morison, R. 2010. Analytics at Work: Smarter Decisions, Better Results, Harvard Business School Press.

Davies, R. and Elias, P. 2003. Dropping Out: A Study of Early Leavers from Higher Education, Research Report RR386, Department for Education and Skills, https://http://www.education.gov.uk/publications/eOrderingDownload/RR386.pdf.

Davis, P. M. and Connolly, M. J. L. 2007. Institutional Repositories Evaluating the Reasons for Non-use of Cornell University's Installation of DSpace, D-Lib Magazine, 13(3/4), http://www.dlib.org/dlib/march07/davis/03davis.html.

Department for Education and Skills 2003. Widening Participation in Higher Education, ref. DfES 0301 2003, https://http://www.education.gov.uk/publications/standard/publicationDetail/Page1/DfES 0301 2003.

DesJardins, S. L. and Pontiff, H. 1999. Tracking institutional leavers: An .pplir.tinn, Tallahassee, Florida: Association for institutional Research.

Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N. and Taibi, D. 2012. Linked education: interlinking educational resources and the web of data. *In proveedings of the 27th ACM Symposium On Applied Computing (SAC-2012), Special Track on Semantic Web and Applications, Trento, Italy*.

Dietze, S., Sanchez-Alonso, S., Ebner, H., Yu, H. Q., Giordano, D., Marenzi, I. and Nunes, B. P. 2013. Interlinking educational Resources and the Web of Data – a Survey of Challenges and Approaches, *Program: electronic library and information systems*, 47(1), 60 - 91.

Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N. and Taibi, D.v. 2012. Linked education: interlinking educational resources and the web of data. *In proveedings of the* 27th ACM Symposium On Applied Computing (SAC-2012), Special Track on Semantic Web and Applications, Trento, Italy.

Department of Innovation, Universities and Skills (DIUS) 2008. Demographic change and its impact on the higher education sector in England, Universities UK, http://webarchive.nationalarchives.gov.uk/+/http://www.bis.gov.uk/wp-content/uploads/2009/10/HE-UUK-demographics.pdf.

Dodgson, R. and Bolam, H. 2002. Student retention, support and widening participation in the North East of England, Universities for the North East, http://www.unis4ne.ac.uk/files/Retention_report70.pdf.

Duda, R. O., Hart, P. E. and Stork, D. G. 2001. *Pattern Classification (2nd edition), Wiley-Interscience.*

Duffy, T. M., Dueber, B. and Hawley, C. L. 1998. *Critical Thinking in a Distributed Environment: A Pedagogical Base for the Design of Conferencing System," in Electronic Collaborators: Learner-Centered Technologies for Literacy, Apprenticeship, and Discourse, Bonk, C. J., King, K. S. (eds.),* Lawrence Erlbaum Associates.

Eaton, J. S. 2009. An overview of US accreditation, Council for Higher Education Accreditation, http://www.chea.org/pdf/2009.06_Overview_of_US_Accreditation.pdf.

Education and Skills Committee 2003. The Future of Higher Education, Fifth Report of Session 2002–03, HC 425-I, vol. 1, London: The Stationery Office Limited, http://www.publications.parliament.uk/pa/cm200203/cmselect/cmeduski/425/425.pdf.

Edward, N. S. 2003. First Impressions Last, *Active Learning in Higher Education,* 4(3), 226-242.

Elkins, S. A., Braxton, J. M. and James, G. W. 2000. Tinto's Separation Stage and its Influence on First-Semester College Student Persistence, *Research in Higher Education,* 41(2).

Feldman, M. J. 1993. Factors associated with one-year retention in a community college, *Research in Higher Education,* 34(41), 503-512.

Field, A. 2009. *Discovering Statistics Using SPSS,* 3rd edition, Sage.

Fox, K. C. 1998. Information Technology in Higher Education: Evolving Learning Environments. *The Higher Education Administrative Technology Conference (CUMREC), Atlanta, Georgia,* http://net.educause.edu/ir/library/pdf/cmr9823.pdf.

Fox, R. 1986. Application of a conceptual model of college withdrawal to disadvantaged students, *American Educational Research Journal,* 23(3), 415-424.

Glynn, J. G., Sauer, P. L., Miller. T. E. 2003. Signaling Student Retention With Prematriculation Data, *NASPA Journal*, 41(1).

Gomez-Perez, A., Fernández-López, M. and Corcho, O. 2004. *Ontological Engineering, Springer, London, UK.*

Greenway, D. and Haynes, M. 2003. Funding Higher Education in the UK: The Role of Fees and Loans, *Economic Journal,* 113(485), F150-F166.

Grosset, J. M. 1991. Patterns of integration, commitment, and student characteristics and retention among younger and older students, *Research in Higher Education,* 32(2), 159-178.

Gruber, T. R. 1993. A translation approach to portable ontology specifications, *Knowledge Acquisition,* 5(2), 199-220.

Guarino, N. 1998. Formal ontology in information systems. *In Proceedings of the 1st International Conference on Formal Ontology in Information Systems (FOIS), Trento, Italy.*

Guerra-López, I. 2008. *Performance Evaluation: Proven Approaches for Improving Program and Organizational Performance, San Francisco, CA94103-1741*, Jossey-Bass. http://www.books.google.ca/books?isbn=1118504895.

Gumport, P. J. and Chun, M. 2005. Technology and higher education: Opportunities and challenges for the new era, Technical report, National Centre for Postsecondary Improvement (NCPI), Stanford, U.S., http://www.immagic.com/eLibrary/ARCHIVES/GENERAL/STANFORD/S000105G.pdf.

Gunn, S. R. 1998. Support Vector Machines for Classification and Regression, Technical Report, ECS, University of Southampton, UK.

Han, J. and Kamber, M. 2000. *Data Mining: Concepts and Techniques, Second edition, Morgan Kaufmann.*

Hanna, D. E. 2003. Building a leadership vision eleven strategic challenges for higher education, *EDUCAUSE review,* http://net.educause.edu/ir/library/pdf/erm0341.pdf.

Hart, M. and Friesner, T. 2004. Plagiarism and Poor Academic Practice – A Threat to the Extension of e-Learning in Higher Education?, *Electronic Journal on e-Learning,* 2(1), 80-96, http://www.business-kac.co.uk/art25.pdf.

Harvey, L., Drew, S. and Smith, M. 2006. The first year experience: a literature review for the Higher Education Academy, York: Higher Education Academy, http://www.heacademy.ac.uk/assets/York/documents/ourwork/research/literature_reviews/first_year_experience_full_report.pdf.

Hausmann, L. R. M., Schofield, J. W. and Woods, R. L. 2007. Sense of belonging as a predictor of intentions to persist among African American and White first-year college students, *Research in Higher Education,* 48(7), 803–839.

Heath, T. and Bizer, C. 2011. *Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology,* First edition, Morgan and Claypool, 1-136.

Higher Education Funding Council for England (HEFCE), 2001. Strategies for widening participation in higher education: A guide to good practice, HEFCE 01/36, available at: http://www.hefce.ac.uk/pubs/hefce/2001/01_36/01_36.pdf.

Higher Education Funding Council for England (HEFCE), 2009a. Future arrangements for quality assurance in England and Northern Ireland, ref: 2009/47, https://http://www.hefce.ac.uk/pubs/year/2009/200947/.

Higher Education Funding Council for England (HEFCE), 2009b. Higher Education in England: Achievements, Challenges and Prospects, ref: 2009/06, http://www.hefce.ac.uk/pubs/year/2009/200906/.

Higher Education Funding Council for England (HEFCE), 2010a. Guide to funding: how HEFCE allocates its funds, ref: 2010/24, http://www.hefce.ac.uk/pubs/hefce/2010/10_24/10_24.pdf.

Higher Education Funding Council for England (HEFCE), 2010b. Investing for successful future: A guide to HEFCE, ref: 2010/23, http://www.hefce.ac.uk/pubs/year/2010/201023/.

Higher Education Funding Council for England (HEFCE), 2014. UK review of the provision of information about higher education: National Student Survey results and trends analysis 2005-2013, ref: 2014/13, http://www.hefce.ac.uk/pubs/year/2014/201413/ - d.en.87641.

Herzog, S. 2005. Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen, *Research in Higher Education,* 46(8), 883-928.

Hilton, T. J. 1982. Persistence in higher education, New York: College Board, (ERIC Document Reproduction Service No. 227 737).

Hirsch, W. Z. and Weber, L. E. 1999. *Challenges Facing Higher Education at the Millennium,* The Oryx Press, http://hdl.handle.net/2027.42/58009.

Ho, W., Higson, H. E., Dey, P. K., Xu, X. and Bahsoon, R. 2009. Measuring performance of virtual learning environment system in higher education, *Quality Assurance in Education,* 17(1), 6-29. http://eprints.aston.ac.uk/551/1/AHPQFD_in_Higher_Education_(QAE).p df.

Hosch, B. 2008. The tension between student persistence and institutional retention:The relationship between first-semester GPA and student progression rates of first-time, full-time students, Association for Institutional Research Annual Forum, Seattle, WA.

Hosmer, D. W. and Lemeshow, S. 2000. *Applied logistic regression,* 2nd edition, New York: Wiley.

Hossler, D. 1984. Emollment management: An intererated apprnarh New York College Entrance Examination Board.

House, D. L. 1994. College grade outcomes and attrition: An exploratory studv of noncognitive variables and academic backgrnnnd as predictors, Shelbyville llflllinois Association for histitutional Research, (ERIC Document Reproduction Service No. ED 241 079).

Hurtado, S. and Carter, D. F. 1997. Effects of college transition and perceptions of the campus racial climate on Latino college students' sense of belonging, *Sociology of Education,* 70 (4), 324–345.

Huynh, D. F., Karger, D. R. and Miller, R. C. 2007. Exhibit: lightweight structured data publishing. *In Proceedings of the 16th international conference on World Wide Web, WWW '07.* New York, USA.

Ishitani, T. 2006. "Studying Attrition and Degree Completion Behavior among First-Generation College Students in the United States." *Journal of Higher Education* 77: 861-885.

Jain, A. K. and Duin, P. W. 2000. Statistical pattern recognition: a review, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22(1), 4-37.

Joint Information Systems Committee (JISC) infoNet 2004. Effective Use of VLEs: Introduction to VLEs, http://www.jiscinfonet.ac.uk/infokits/vle/.

Joint Information Systems Committee (JISC), 2007. Student Expectations Study, http://www.jisc.ac.uk/media/documents/publications/studentexpectations.pdf.

Joint Information Systems Committee (JISC), 2008. Great expectations of ICT: How Higher Education institutions are measuring up, http://www.jisc.ac.uk/media/documents/publications/jiscgreatexpectationsfinalreportjune08.pdf.

Joint Information Systems Committee (JISC), 2009a. Communicating knowledge: How and why UK researchers publish and disseminate their findings, http://www.jisc.ac.uk/publications/research/2009/communicatingknowledgereport.

Joint Information Systems Committee (JISC), 2009b. JISC Strategy 2010–2012, http://www.jisc.ac.uk/publications/strategy/2009/strategy2010.aspx.

Joint Information Systems Committee (JISC), 2010. Strategic engagement with business and the community, http://www.jisc.ac.uk/publications/briefingpapers/2010/bpstrategicengagement.aspx.

Johnson, L., Smith, R., Willis, H., Levine, A. and Haywood, K. 2011. The 2011 Horizon Report. Austin, Texas: The New Media Consortium, http://net.educause.edu/ir/library/pdf/hr2011.pdf.

Johnson, M. M. and Molnar, D. 1996. Comparing retention factors for anglo, black and hispanic students, Albuquerque, NM: Association for histitiitional Research.

Johnston, V. 2001. Developing Strategies to Improve Student Retention: Reflections from the Work of Napier University's Student Retention Project. SRHE conference. Cambridge.

Jones, D. R. 2008. Student retention and Success - synthesis of research literature in the area of retention, Higher Education Academy, http://www.heacademy.ac.uk/assets/EvidenceNet/Syntheses/wp_retention_synthesis_for_pdf_updated_090310.pdf.

Kalfoglou, Y. and Schorlemmer, M. 2003. Ontology mapping: the state of the art, *The knowledge engineering review,* 18(1), 1-31.

Kam, Adele H.T. and Ch'ng, P. C. 2009. Predicting Academic Performance of Engineering Diploma Students, *In the proceedings of the 2nd International Conference of Teaching and Learning (ICTL 2009),* INTI University College, Malaysia.

Kay, D. and van Harmelen, M. 2012. CETIS Analytics Series, Vol. 1, No. 2, Analytics for the Whole Institution: Balancing Strategy and Tactics, http://publications.cetis.ac.uk/wp-content/uploads/2012/11/Analytics-for-the-Whole-Institution-Vol1-No2.pdf.

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

Kiefer, C., Bernstein, A. and Locher, A. 2008. Adding Data Mining Support to SPARQL via Statistical Relational Learning Methods, *European Semantic Web Conference,* 478-492.

Kim, J. 2007. Motivating and Impeding Factors Affecting Faculty Contribution to Institutional Repositories, *Journal of Digital Information,* 8(2), http://journals.tdl.org/jodi/index.php/jodi/article/view/193/177.

Klump, J., Wa̎chter, J. and the STD-DOI Consortium 2004. Open access to data and the 'Berlin Declaration, The International CODATA Conference. Berlin, Germany, http://www.codata.org/04conf/papers/Klump-paper.pdf.

Kotsiantis, S., Pierrakeas, C. and Pintelas, P. 2004. Predicting Students' Performance in Distance Learning Using Machine Learning Techniques, *Applied Artificial Intelligence (AAI)*, 18(5), 411-426.

Kovacic, Z. J. 2010. Early prediction of student success:Mining student enrollment data, *In the proceedings of Informing Science and IT Education Conference (InSITE) 2010*. http://proceedings.informingscience.org/InSITE2010/InSITE10p647-665Kovacic873.pdf

Kovacic, Z. J. and Green, J. S. 2010. Predictive working tool for early identification of 'at risk' students, Open Polytechnic, New Zealand. http://akoaotearoa.ac.nz/download/ng/file/group-6/predictive-working-tool-for-early-identification-of-at-risk-students---full-report.pdf

Kuh, G. D. 2009. What Student Affairs Professionals Need to Know about Student Engagement, *Journal of College Student Development,* 50(6), 683-706.

Langbein, L. I. and Snider, K. 1999. The impact of teaching on retention: Some quantitative evidence, *Social Science Quarterly,* 80(3), 457–472.

Lauría, E. J. M., Baron, J. D., Devireddy, M., Sundararaju, V. and Jayaprakash, S. M., 2012. Mining academic data to improve college student retention: An open source perspective. *In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 139-142, Vancouver, Canada.

Lenning, O. T., Sauer, K. and Beal, P. E. 1980. Student retention strategies, Washington, D.C: American Association for Higher Education.

Li, G. a. and Kilhan, T. 1999. Students who left college: An examination of their characteristics and reasons for leaving. Seattle, WA: Association for Institutional Research. (ERIC Document Reproduction Service No. 433 779).

Light, A., and Strayer, W. 2000. Determinants of college completion, *Journal of Human Resources,* 35(2), 299-332.

Linting, M., Meulman, J. J., Groenen, P. J. F. and Van der Kooij, A. J. 2007. Nonlinear Principal Components Analysis: Introduction and Application., *Psychological Methods,* 12(3), 336–358. http://psych.colorado.edu/~willcutt/pdfs/Linting_2007.pdf

Long, P. and Siemens, G. 2011. Penetrating the Fog: Analytics in Learning and Education, EDUCAUSE review, 46 (5), https://net.educause.edu/ir/library/pdf/ERM1151.pdf.

Looney, C. G. 1997. *Pattern Recognition Using Neural Networks,* Oxford University Press.

Lowe, H. and Cook, A. 2003. Mind the Gap: are Students Prepared for Higher Education? , *Journal of Further and Higher Education* 27(1), 53-76.

Lynch, C. A. 2003. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age, http://scholarship.utm.edu/21/1/Lynch,_IRs.pdf.

McAuley, D., Rahemtulla, H., Goulding, J. and Souch, C. 2011. How Open Data, data literacy and Linked Data will revolutionise higher education, http://pearsonblueskies.com/2011/how-open-data-data-literacy-and-linked-data-will-revolutionise-higher-education/.

McCord, A. 2003. Institutional repositories: Enhancing teaching, learning, and research, *EDUCAUSE National Conference, https://net.educause.edu/ir/library/pdf/dec0303.pdf.*

Milem, J. and Berger, J. 1997. A modified model of college student persistence: Exploring the relationship between Astin's theory of involvement and Tinto's theory of student departure, *Journal of College Student Development,* 38(4), 387-400.

Miller, P. 2010. Linked Data Horizon Scan, Joint Information Systems Committee (JISC), http://linkeddata.jiscpress.org/.

Miller, T. E. and Herreid, C. H. 2008. Analysis of Variables to Predict First-Year Persistence Using Logistic Regression Analysis at the University of South Florida, *College and University,* 83(3), 2-11.

Miller, T. E. and Tyree, T. M. 2009. Using a model that predicts individual student attrition to intervene with those who are most at risk, *College and University,* 84(3), 12-19.

Milligan, C. 1999. Delivering Staff and Professional Development Using Virtual Learning Environments, http://www.jisc.ac.uk/media/documents/programmes/jtap/jtap-044.pdf.

Mitra, P. and Wiederhold, G. 2002. Resolving terminological heterogeneity in ontologies. *Proceedings of the ECAI'02 workshop on Ontologies and Semantic Interoperability.*

Mitsopoulou, E., Taibi, D., Giordano, D., Dietze, S., Yu, H. Q., Bamidis, P., Bratsas, C. and Woodham, L. 2011. Connecting Medical Educational Resources to the Linked Data Cloud: the mEducator RDF Schema, Store and API, *In: Linked Learning 2011, Proceedings of the 1st International Workshop on eLearning Approaches for the Linked Data Age, 8th extented semantic web conference (ESWC2011), Heraklion, Greece. http://oro.open.ac.uk/29124/1/paper4.pdf.*

Morrison, H. G. 2006. Dramatic Growth of Open Access: Implications and Opportunities for Resource Sharing, *Journal of Interlibrary Loan, Document Delivery and Electronic Reserve,* 16(3), http://eprints.rclis.org/6680/6681/dramatic.pdf.

Moore, S. A., D'Addario, D. M., Kurinskas, J. and Weiss, G. M. 2009. Are Decision Trees Always Greener on the Open (Source) Side of the Fence?, *In Proceedings of the 2009 International Conference on Data Mining*, *CSREA Press,* 185-188.

Moxley, D., Najor-Durack, A. and Dumbrigue, C. 2001. *Keeping Students in Higher Education: Successful practices and strategies for retention, London*, Kogan Page.

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

National Audit Office (NAO), 2002. Improving student achievement in English higher education, London: The Stationery Office, http://www.nao.org.uk/publications/0102/improving_student_achievement.aspx.

National Audit Office (NAO), 2007. Staying the course: the retention of students in higher education, London: The Stationery Office, http://www.nao.org.uk/publications/0607/retention_of_students_in_he.aspx.

Napoli, A. R. and Wortman, P. M. 1996. A meta-analysis of the impact of academic and social integration on persistence of community college students, *Journal of Applied Research in the Community College,* 4, 5–21.

Natsu, J. 2010. Advanced Analytics: Helping Educators Approach the Ideal, eSN Special Report, eSchool News, 17–23, http://www-935.ibm.com/services/ie/gbs/pdf/Smarter_Education_Advanced_Analytics.pdf.

Noel, L. 1985. *Incresing student retention: New challenges and potential. In L. Noel, R. Levitz, and D. Saluri (Eds.), Increasing student retention (1-27),* San Francisco: Jossey-Bass.

Nora, A., Attinasi, L. C. and Matonak, A. 1990. Testing qualitative indicators of precollege factors in Tinto's attrition model: A community college student population, *The Review of Higher Education,* 13(3), 337-356.

Organization of Economic Co-operation and Development (OECD), 2013. Education at a Glance 2013: OECD Indicators, OECD Publishing, http://dx.doi.org/10.1787/eag-2013-en.

O'Leary, R. 2002. Virtual Learning Environments, Association for Learning Technology, http://www.alt.ac.uk/docs/eln002.pdf.

Omitola, T., Koumenides, C. L., Popov, I. O., Yang, Y., Salvadores, M., Correndo, G., Hall, W. and Shadbolt, N. 2010. Integrating Public Datasets Using Linked Data: Challenges and Design Principles Future Internet Assembly, *Future Internet Assembly,* Ghent, Belgium.

Oregon State System of Higher Education 1994. The long and winding road: Retention, attrition, and graduation of OSSHE freshmen entering 1986-87, Eugene, OR: Oregon State System of Higher Education.

Ounnas, A., Liccardi, I., Davis, H., Millard, D. E. and White, S. A. 2006. Towards a Semantic Modeling of Learners for Social Networks, International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL) at the AH2006 Conference, Dublin, Ireland. pp. 102-108.

Parliamentary Select Committee on Education and Employment 2001. Higher Education: Student Retention, Sixth Report, London: The Stationery Office.

Pascarella, E. T. 1980. Student-faculty informal contact and college outcomes, *Review of Educational Research,* 50, 545-595.

Pascarella, E. T. and Terenzini, P. T. 1980. Predicting freshman persistence and voluntary dropout decisions from a theoretical model, *Journal of Higher Education,* 51(1), 60–75.

Pascarella, E. T., Duby, P. B. and Iverson, B. K. 1983. A Test and Reconceptualization of a Theoretical Model of College Withdrawal in a Commuter Institution Setting, *Sociology of Education,* 56(2), 88-100.

Pascarella, E. T., Terenzini, P. T. and Wolfle, L. M. 1986. Orientation to college and freshman year persistence/withdrawal decisions, *Journal of Higher Education,* 57(2), 155–175.

Pascarella, E. T. and Terenzini, P. T. 2005. How College Affects Students: A third decade of research, *Journal of College Student Development,* 47(5), 589-592.

Peng, C. J., So, T. H., Stage, F. K. and St. John, E. P. 2002. The use and interpretation of logistic regression in higher education journals: 1988–99, *Research in Higher Education,* 43(3), 259–293.

Porter, O. F. 1990. Undergraduate completion and persistence at four-year colleges and universities. Washington, D.C: National Institute of Independent Colleges and Universities.

Price, L. A. 1993. Undergraduate completion and persistence at four-year colleges and universities. Cumberland, MD: Allegany Community College. (ERIC Document Reproduction Service No. 361 051).

Prime Minister's Strategy Unit 2007. Higher Education: Progress, challenges and a new scheme to promote voluntary giving, http://dera.ioe.ac.uk/id/eprint/7068.

Quality Assurance Agency (QAA), 2008a. Outcomes from institutional audit: Collaborative provision in the institutional audit reports, second series, http://dera.ioe.ac.uk/1757/.

Quality Assurance Agency (QAA), 2008b. Outcomes from institutional audit: Learning support resources (including virtual learning environments), Second series, http://www.qaa.ac.uk/Publications/InformationAndGuidance/Pages/Outcomes-from-institutional-audit-Second-series-Learning-support-resources-including-virtual-learning-environments.aspx.

Quality Assurance Agency (QAA), 2008c. Outcomes from institutional audit: Progression and completion statistics, Second series, http://www.qaa.ac.uk/reviews/institutionalAudit/outcomes/series2/Institutions progressioncompletion.pdf.

Quality Assurance Agency (QAA), 2008d. Outcomes from institutional audit: Staff support and development arrangements, Second series http://www.qaa.ac.uk/Publications/InformationAndGuidance/Pages/Outcomes-from-institutional-audit---Staff-support-and-development---Second-series.aspx.

Quality Assurance Agency (QAA), 2008e. Outcomes from institutional audit: Work-based and placement learning, and employability, Second series, http://dera.ioe.ac.uk/9644/.

Quality Assurance Agency (QAA), 2010. Information bulletin: Integrated quality and enhancement review - Assessment and Review.

Quality Assurance Agency (QAA), 2012. UK Quality Code for Higher Education - Chapter B3: Learning and Teaching, http://www.qaa.ac.uk/en/Publications/Documents/quality-code-B7.pdf.

Quadri, M. N. and Kalyankar, N. V. 2010. Drop Out Feature of Student Data for Academic Performance Using Decision Tree., *Global Journal of Computer Science and Technology,* 10(2).

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning,* Morgan Kaufmann Publishers, San Mateo, CA.

Rae, D. 2007. Connecting enterprise and graduate employability: Challenges to the higher education culture and curriculum?, *Education + Training,* 49(8/9), 605-619, http://www.emeraldinsight.com/journals.htm?articleid=1637369.

Ramist, L. 1981. College student attrition and retention, New York: College Board, (ERIC Document Reproduction Service No. ED 200 170).

Randall, M. 1999. Retention and graduation rates at Maryland four-year public institiitions, Annapolis, MD: Maryland State Higher Education Commission, (ERIC Document Reproduction Service No. ED 432 165).

Raym, C. 2006. The Case for Institutional Repositories: A SPARC Position Paper. Discussion Paper, Scholarly Publication and Academic Resources Coalition, Washington, D.C., http://scholarship.utm.edu/20/.

Reinsel, D., Chute, C., Schlichting, W., Mcarthur, J., Minton, S., Xheneti, I., Toncheva, A. and Manfrediz, A. 2007. The Expanding Digital Universe, An IDC White Paper, http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf

Rendon, L., Jalomo, R. and Nora, A. 2000. *Theoretical considerations in the study of minority student retention in higher education. In J. Braxton (Ed.), Reworking the student departure puzzle (127-153).* Nashville: Vanderbilt University Press.

Robst, J., Keil, J. and Russo, D. 1998. The effect of fender composition of faculty on student retention, *Economics of Education Review,* 17(4), 429–439.

Roweton, W. E. 1994. Predicting rural r.ollege retention .mpung first-year undergraduates (ERIC Document Reproduction Service No. 370 501).

Sarker, F., Davis, H. and Tiropanis, T. 2010a. A Review of Higher Education Challenges and Data Infrastructure Responses, *In the proceedings of 3rd International Conference of Education Research and Innovation (ICERI2010),* Madrid, Spain, 1473-1483, http://eprints.soton.ac.uk/271695/1/A_Review_of_HE_challenges_and_data_infrastructure_responses.pdf.

Sarker, F., Davis, H. and Tiropanis, T. 2010b. The Role of Institutional Repositories in addressing Higher Education Challenges, *In the proceedings of 2nd International Workshop on Semantic Web Applications in Higher Education (SEMHE 2010)*, University of Southampton, Southampton, UK, http://eprints.soton.ac.uk/271694/1/The_Role_of_Institutional_Repositories_in_addressing_Higher_Education_Challenges.pdf.

Sarker, F., Tiropanis, T. and Davis, H. 2013a. Exploring Student Predictive Model that Relies on Institutional Databases and Open Data Instead of Traditional Questionnaires, *In the proceedings of 3rd International Workshop on Learning and Education with the Web of Data (LILE2013), WWW'13 companion,* Rio de Janeiro, Brazil, 413-418, http://www2013.wwwconference.org/companion/p413.pdf.

Sarker, F., Tiropanis, T. and Davis, H. 2013b. Students' Performance Prediction by Using Institutional Internal and External Open Data Sources, *In the proceeding of 5th International Conference on Computer Supported Education (CSEDU2013).* Aachen, Germany, 8, http://eprints.soton.ac.uk/353532/1/Students'%20mark%20prediction%20model.pdf.

Satya, S. S., Wolfgang, H., Sebastian, H., Kingsley, I., Ted, T. J., Sören, A., Juan, S. and Ahmed, E. 2009. A Survey of Current Approaches for Mapping of Relational Databases to RDF, http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf.

Schwartz, C., Barron, M. L. and Mauger, A. J. 2010. Using Technology to Impact Student Retention at Montgomery County Community College, EDUCAUSE Quarterly Magazine, 33(4).

Sehgal, L., Mohan, N. and S. Sandhu, P. 2012. Quality Prediction of Function Based Software Using Decision Tree Approach, International Conference on Computer Engineering and Multimedia Technologies (ICCEMT'2012) September 8-9, 2012, Bangkok (Thailand).

Seidman, A. 2005. *College student retention: Formula for Success, Westport*, Praeger Publishers, http://books.google.co.uk/books?id=ckk5B_ADM_YC&printsec=frontcover&source=gbs_ge_summary_r&cad=0 - v=onepage&q&f=false.

Sembiring, S., Zarlis, M., Hartama, D., Ramliana S and Wani, E. 2011. Prediction of student academic performance by an application of data mining techniques, *In the International Proceedings of Economics Development and Research (IPEDR),* 6, http://www.ipedr.com/vol6/21-A10015.pdf.

Shadbolt, N., Berners-Lee, T. and Hall, W. 2006. The semantic web revisited, *IEEE Intelligent Systems,* 21, 96-101.

Siemens, G. and Baker, R. 2012. Learning analytics and educational data mining: towards communication and collaboration. *In the Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, Canada, 1–3.*

Singell, L. D. and Waddell, G. R. 2010. Modeling Retention at a Large Public University: Can At-Risk Students Be Identified Early Enough to Treat?, *Research in High Education,* 51, 546-572.

Spady, W. G. 1970. Dropouts from higher education:An interdisciplinary review and synthesis, *Interchange,* 1(1), 64-85.

Spiller, D. 2009. Assessment: Feedback to Promote Student Learning, Teaching Development, The university of Waikato, http://www.waikato.ac.nz/tdu/pdf/booklets/6_AssessmentFeedback.pdf

Stage, F. K. 1988. University attrition: LISREL with logistic regression for the persistence criterion, *Research in Higher Education,* 29(4), 343-357.

Sullivan, A. V. S. 1997. Rites and passages: students' views of academic and social integration, *College Student Affairs Journal,* 16(2), 4–14.

Terenzini, P. T. and Pascarella, E. T. 1978. The relation of students' precollege characteristics and freshman year experience to voluntary attrition, *Research in Higher Education,* 9, 347-366.

Terenzini, P. T. and Pascarella, E. T. 1980. Toward the Validation of Tinto's model of College Student Retention: A review of recent studies, *Research in Higher Education,* 12(3), 271-282.

Terenzini, P. T., Lorang, W. G. and Pascarella, E. T. 1981. Predicting Freshman Persistence and voluntary dropout decisions:A replication, *Research in Higher Education,* 15(2), 109-127.

Terenzeni, P. T., Pascarella, E. T., Theophilides, C. and Lorang, W. G. 1985. A replication of a path analytic validation of Tinto's theory of college stiident attiition, *Review of Higher Education,* 8(4), 319-340.

The House of Commons 2008. Staying the course:the retention of students on higher education courses, London: The Stationery Office Limited, http://www.publications.parliament.uk/pa/cm200708/cmselect/cmpubacc/322/322.pdf

The House of Commons 2008. Staying the course: the retention of students on higher education courses, London: The Stationery Office Limited.

The House of Commons 2013. Higher education finance statistics, SN/SG/5440, http://www.parliament.uk/briefing-papers/SN05440.pdf.

The sub-committee for Teaching Quality and the Student Experience 2009. HEFCE's statutory responsibility for quality assurance, ref:2009/40, http://www.hefce.ac.uk/pubs/year/2009/200940/.

Thomas, L. 2002. Student retention in higher education: The role of institutional habitus, *Journal of Educational Policy,* 17(4), 423-442.

Thomas, M., Adams, S. and Birchenough, A. 1996. Student Withdrawal from Higher Education, *Educational Management and Administration,* 24(2), 207-221.

Thomas, S. 2000. Ties that bind: A social network approach to understanding student integration and persistence, *Journal of Higher Education,* 71(5), 591-615.

Tinto, V. 1975. Dropout from Higher Education: A Theoretical Synthesis of Recent Research, *Review of Educational Research,* 45(1), 89-125.

Tinto, V. 1987. *Leaving Early: Rethinking the Causes and Cures of Student Attrition,* Chicago: The University of Chicago Press.

Tinto, V. 1988. Stages of student departure: Reflections on the longitudinal character of student leaving, *Journal of Higher Education,* 59(4), 438-455.

Tinto, V. 1993. *Leaving College: rethinking the Causes and Cures of Student Attrition, Chicago*, 2nd edition, Chicago: The University of Chicago Press.

Tinto, V. 1997. Classrooms as communities: Exploring the educational character of student persistence, *Journal of Higher Education,* 68(6), 599-623, http://jesserbishop.wiki.westga.edu/file/view/Tinto_Classrooms+as+Communities.pdf.

Tinto, V. 1998. Colleges as communities: taking research on student persistence seriously, *Review of Higher Education,* 21(2), 167-177.

Tinto, V. 2000. Taking Student Retention Seriously: Rethinking the First Year of College, *NACADA Journal,* 19(2), 5-10.

Tinto, V. 2006-2007. Research and Practice of Student Retention: What Next?, *Journal of College Student Retention*, 8(1), 1-19.

Tiropanis, T., Davis, H., Millard, D. and Weal, M. 2009a. Semantic Technologies for Learning and Teaching in the Web 2.0 Era, *Intelligent Systems, IEEE,* 24(6), 49-53, http://eprints.soton.ac.uk/268429/.

Tiropanis, T., Davis, H., Millard, D. and Weal, M. 2009b. Semantic Technologies for Learning and Teaching in the Web 2.0 era - A survey, In: *WebSci'09: Society On-Line.* Athens, Greece, http://eprints.soton.ac.uk/267106/.

Tiropanis, T., Davis, H., Millard, D., Weal, M. and White, S. 2009c. Linked Data as a Foundation for the Deployment of Semantic Application in Higher Education, In, *SWEL'09: Ontologies and Social Semantic Web for Intelligent Educational Systems, AIED'09,* Brighton, UK, http://eprints.soton.ac.uk/267660/.

Tiropanis, T., Davis, H., Millard, D., Weal, M., White, S. and Wills, G. 2009d. Semantic Technologies in Learning and Teaching (SemTech), JISC Technical Report, http://eprints.ecs.soton.ac.uk/17534/.

Trowler, V. 2010. Student engagement literature review, The higher education academy, http://www.heacademy.ac.uk/assets/documents/studentengagement/StudentEngagementLiteratureReview.pdf.

Universities UK, 2012a. Futures for Higher Education: Analysing Trends. Higher Education: Meeting the Challenges of the 21st Century, http://www.universitiesuk.ac.uk/highereducation/Documents/2012/FuturesForHigherEducation.pdf.

Universities UK, 2012b. Patterns and Trends in the UK Higher Education, http://www.universitiesuk.ac.uk/highereducation/Documents/2012/PatternsAndTrendsinUKHigherEducation2012.pdf.

Universities UK, 2013. Patterns and Trends in the UK Higher Education, http://www.universitiesuk.ac.uk/highereducation/Documents/2013/PatternsAndTrendsinUKHigherEducation2013.pdf.

Universities UK, 2014. The Funding Environment for Universities 2014: Research and Postgraduate Research Training, http://www.universitiesuk.ac.uk/highereducation/Documents/2014/ResearchAndPGRtraining.pdf

van Barneveld, A., Arnold, K. E. and Campbell, J. P. 2012. Analytics in Higher Education: Establishing a Common Language, ELI Paper 1, EDUCAUSE, http://net.educause.edu/ir/library/pdf/ELI3026.pdf.

van Harmelen, M. and Workman, D. 2012. CETIS Analytics Series, Vol. 1, No. 3, Analytics for Learning and Teaching, http://publications.cetis.ac.uk/2012/516.

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

Waggener, A. T. a. and Smith, C. K. 1993. Benchmark factors in student retention, New Orleans, LA: Mid-South Educational Research Association.

Walker, L. 1999. Longitudinal study of drop-out and continuing students who attended the Pre-University Summer School at the University of Glasgow, *International Journal of Lifelong Education,* 18(3), 217-233.

Wall, M. 1996. From theory to practice: Using retention research to guide assessment efforts at a community college . Mays Landing, NJ: Atlantic Community College. (ERIC Document Reproduction Service No. ED 397 929).

West, G. B. 1999. Teaching and Technology in Higher Education: Changes and Challenges, *Journal article of Adult Learning,* 10, 16, http://www.elcamino.edu/faculty/jsuarez/11Cour/H10A/ECCSTWest.pdf

Whitaker, D. G. 1987. Persistence and the two-year college student. Baltimore, MD: Association for the Study of Higher Education. (ERIC Document Reproduction Service No. 292 404).

World Wide Web Consortium. 2004. *RDF primer. Recommendation, W3C,* http://www.w3.org/TR/rdf-primer/.

World Wide Web Consortium. 2008. *SPARQL query language for RDF. Recommendation, W3C, URL:* http://www.w3.org/TR/rdf-sparql-query/.

Wright, R. E. 1995. *Logistic Regression*, L.G. Grimm and P.R. Yamold (Eds.), *Reading and Understanding Multivariate Statistics* (217-243), Washington, D.C: American Psychological Association.

Yadav, S. K., Bharadwaj, B. K. and Pal, S. 2011. Data Mining Applications: A comparative study for predicting students' performance, *International journal of Innovative Technology and Creative Engineering (IJITCE),* 1(12).

Yadev, S. K. and Pal, S. 2012. Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification, *World of Computer Science and Information Technology (WCSIT),* 2(2), 51-56.

York, M. C. 1993. Causes of college retention: A systems perspective, Toronto, Ontario: American Psychological Association.

Yorke, M. 1999. *Leaving Early: Undergraduate Non-completion in Higher Education*, London: Falmer Press.

Yorke, M. and Longden, B. 2004. *Retention and student success in higher education, Maidenhead*, SRHE and Open University Press.

Yorke, M. and Thomas, L. 2003. Improving the retention of students from lower socio-economic groups, *Journal of Higher Education Policy and Management,* 25(1), 63-75.

Zapilko, B. and Mathiak, B. 2011. Performing statistical methods on linked data. *International conference on dublin core and metadata applications.*

Zhang, Z. A. and RiCharde, R. S. 1998. Prediction and analysis of Freshman retention, Mirmeapolis, MN: Association for Institutional Research.

Zhao, C. and Luan, J. 2006. Data mining: Going beyond traditional statistics, *New Directions for Institutional Research, 131(2)*, 7-16.

Zheng, J. L., Saunders, K. P., Shelley, M. C. and Whalen, D. F. 2002. Predictors of academic success for freshmen residence hall students, *Journal of College Student Development,* 43(2), 267-283.

# Appendix A

## Questionnaires

This online-questionnaire was used to collect student data, which have been traditionally used in developing student predictive model to predict at-risk students based on Tinto's student retention model. It comprises of 49 questions in six sections. The questions are listed in the following.

## Section 1

1. **What is your study commencement date** (if forget any tentative date you can remember)**?**

My answer is:

2. **What is the length of the programme of your study?**

My answer is:

3. **What is the title of the programme you have been enrolled (e.g. BSc in CSC, etc.)?**

My answer is:

4. **What is your current year of study?**

a) 1st b) 2nd c) 3rd

I. **In which semester?**

a) 1st b) 2nd

5. **What is your mode of study?**

a) Full time b) Part time

6. **What was your age when you enrolled in your programme of study?**

My answer is:

7. **What is your gender?**

a) Male b) Female

8. **What is your ethnicity (e.g. White, Black, Asian, etc.)?**

My answer is:

**9. What is the name of your home country?**

My answer is:

**10. Is your Residence?**

a) UK b) Other-EU c) Non-EU

**11. Do you have any disability (registered as disabled with the university)?**

a) Yes b) No

**12. What is your marital status?**

a) Married b) Unmarried

**13. What was your employment status when you were in year 1 of your study?**

a) Employed b) Unemployed

*If your answer is **employed**, please proceed to question number 14 and if your answer is **unemployed** proceed to question number 17.

**14. Was it a full time/part time employment?**

a) Full time b) Part time

**15. How many hours you worked per week?**

My answer is:

**16. Where did you work (on campus/off campus)?**

a) On campus b) Off campus

**17. What was your qualification on entry in your programme of study?**

a) A levels
b) Foundation course HE or FE level
c) Baccalaureate
d) HE degree
e) Other qualification, please specify here:

**18. Did the grades of your previous qualification match with the university requirement/ offer?**

a) Yes b) No

**19. What was your tariff point?**

a) 0-100 b) 101-160 c) 161-200 d) 201-230 e) 231-260
f) 261-290 g) greater than 290

**20. What is your previous school type?**

a) State school
b) Independent school
c) Other, please specify here:

**21. Did you have any previous higher education experience before entry on your current programme?**

a) Yes b) No

**22. How do you know about this institution?**

a) Open day
b) University Tour
c) Online
d) Paper prospectus
e) Talking to friends
f) Others
g) Not applicable

**23. How do you know about your programme of study?**

a) Open day
b) University Tour
c) Online
d) Paper prospectus
e) Talking to friends
f) Others
g) Not applicable

**24. Are you first in your family to enter in higher education?**

a) Yes b) No

**25. What is your (parent, step-parent or guardian) socio-economic group/status?**

a) Higher managerial and professional occupations
b) Lower managerial and professional occupations
c) Intermediate occupations
d) Small employers and own account workers
e) Lower supervisory and technical occupations
f) Semi-routine occupations
g) Routine occupations
h) Not classified / unknown

**26. What is your fee status?**

a) Home b) Overseas

**27. What is your source of fee?**

a) Yourself b) Family c) Grant/Scholarship d) Student loan e) Others

**28. How were you admitted to this institution?**

a) Direct b) Clearing c) Others, please specify here:

**29. Do you have any responsibility for dependants (e.g. spouse, children, etc.)?**

a) Yes b) No

**30. Where did you live during the term-time in your 1st year?**

a) University halls
b) Private halls
c) Parents house
d) Own residence
e) Rented accommodation
f) Other

31. **What is your mother's educational level?**

a) Higher Education (e.g. BSc, MSc). b) Further education (e.g. A levels).
c) GCSE d) Did not go to school e) Do not know

32. **What is your father's educational level?**

a) Higher Education (e.g. BSc, MSc). b) Further education (e.g. A levels).
c) GCSE d) Did not go to school e) Do not know

33. **What is your mother's occupation?**

My answer is:

34. **What is your father's occupation?**

My answer is:

35. **What is your father's annual income (£)?**

a) <= 15,000 b) 15,001 - 20,000 c) 20,001 - 25,000 d) 25,001 – 30,000 e) 30,001 – 35,000 f) 35,001 – 40,000 g) 40,001- 50,000 h) >50,001

36. **What is your mother's annual income(£)?**

a) <= 15,000 b) 15,001 - 20,000 c) 20,001 - 25,000 d) 25,001 – 30,000 e) 30,001 – 35,000 f) 35,001 – 40,000 g) 40,001- 50,000 h) >50,001

37. **What was the term time postal code during your 1st year of study?**

My answer is:

38. **What is your parental/permanent postal code in UK?**

My answer is:

39. **How much time do you need to travel from your term time residence to your university (in hours)?**

My answer is:

40. **What was your 1st year Marks (%)?**

a) 70-100 b) 60-70 c) 50-60 d) 40-50 e) Fail

I. **What was your 1st semester Marks (%)?**

a) 70-100 b) 60-70 c) 50-60 d) 40-50 e) Fail

II. **What was your 2nd semester Marks (%)?**

a) 70-100 b) 60-70 c) 50-60 d) 40-50 e) Fail

41. **Did you attend your department/course fresher induction event(s)?**

a) Yes b) No

42. **Did you attend any education opportunities/training in the university beyond your course (e.g. cv skills, writing skills, workshops, etc.)?**

a) Yes b) No

**43. Your choice of this institution was**

a) 1st b) 2nd c) lower than 2nd d) Not applicable/clearing.

**44. Did you change your tutor during the 1st year of your study? If yes, in which semester you have changed? If no, please proceed to question number 45.**

a) 1st semester b) 2nd semester

## Section 2

**45. Peer group interactions (This question concerns your interaction with your classmates or other students).**

**I. Since coming to this university, I have developed close personal relationships with other students.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**II. The student friendships I have developed at the university have been personally satisfying.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**III. My interpersonal relationships with other students have had a positive influence on my personal growth, attitudes and values.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**IV. My interpersonal relationships with other students have had a positive influence on my intellectual growth and interest in ideas.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**V. It has been difficult for me to meet and make friends with other students.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**VI. Few of the students I know would be willing to listen to me and help me if I had a personal problem.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**VII. Most students at this university have values and attitudes different from my own.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

## Section 3

**46. Interactions with faculty staff (This question concerns your interaction with your faculty staff outside of lectures (e.g. tutorials, meetings, chats, etc.))**

**I. My non-classroom interactions with faculty have had a positive influence on my personal growth, values, and attitudes.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**II. My non-classroom interactions with my faculty have had a positive influence on my intellectual growth and interest in ideas.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**III. My non-classroom interactions with faculty have had a positive influence on my career goals and aspirations.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**IV. Since coming to this university, I have developed a close, personal relationship with at least one faculty member.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**V. I am satisfies with the opportunities to meet and interact informally with faculty members.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

## <u>Section 4</u>

**47. Faculty Concern for Student Development and Teaching.**

**I. Few of the faculty members I have had contact with are generally interested in students.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**II. Few of the faculty members I have had contact with are generally outstanding or superior teachers.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**III. Few of the faculty members I have had contact with are willing to spend time outside of class to discuss issue of interest and importance to students.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**IV. Most of the faculty I have had contact with are interested in helping students grow in more than just academic areas.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**V. Most faculty members I have had contact with are genuinely interested in teaching.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**VI. I am satisfied with the support my tutor provided me during my 1st year?**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

## <u>Section 5</u>

**48. Academic and Intellectual Development**

**I. I am satisfied with the extent of my intellectual development since enrolling in this university.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**II. My academic experience has had a positive influence on my intellectual growth and interest in ideas.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**III. I am satisfied with my academic experience at this university.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**IV. Few of my coursers this year have been intellectually stimulating.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**V. My interest in ideas and intellectual matters has increased since coming to this university.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**VI. I am more likely to attend a cultural event (for example, a concert, lecture, or art show) now than I was before coming to this university.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**VII. My academic performance has met my expectations.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

## Section 6

**49. Institutional and Goal Commitment**

**I. It is important for me to graduate from this university.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**II. I am confident that I made the right decision in choosing to attend this university.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**III. It is likely that I will register at this university next year.**

a) Strongly agree b) Agree c) Disagree d) Strongly disagree

**IV. Getting good grades is not important to me.**

a) Strongly agree b) Agree c) Disagree d) Strongly Disagree

# Appendix B

## Ethics Review Documentation

Ethics ID: 1978

The following is the application, which was submitted to get the ethics approval from the ethics committee of the university of Southampton. Initially the survey duration was for one month 10 days, later we extended to get our expected number of participants.

## ERGO application form – Ethics form

**All mandatory fields are marked (M*). Applications without mandatory fields completed are likely to be rejected by reviewers. Other fields are marked “if applicable”. Help text is provided, where appropriate, in italics after each question.**

### 1. Applicant Details

| | |
|---|---|
| **1.1 (M*) Applicant name:** | |
| **1.2 Supervisor (if applicable):** | |
| **1.3 Other researchers/collaborators (if applicable):** *Name, address, email, telephone* | |

### 2. Study Details

| | |
|---|---|
| **2.1 (M*) Title of study:** | |
| **2.2 (M*) Type of study** (*e.g. Undergraduate, Doctorate, Masters, Staff*): | Doctorate |

| | |
|---|---|
| **2.3 i) (M*) Proposed start date:** | 20/05/2012 |
| **2.3 ii) (M*) Proposed end date:** | 30/06/2012 |

**2.4 (M*) What are the aims and objectives of this study?**

The aim of this study is to examine the capability of linked data technologies and the sufficiency of existing linked data repositories in supporting student retention, progression and completion. A further aim is to build a student predictive model through integrating linked data sources that are internal or external to higher education institutions, without having to rely on questionnaire data that have been essential in existing models.

**2.5 (M*) Background to study** (*a **brief** rationale for conducting the study*):

Student retention, progression and completion has proven one of the key issues to be addressed by the higher education institutions around the world. Research on student retention, progression and completion is traditionally survey-based. The major problem/limitation with survey-based study is the lower participation rate, which affects the study output. We propose ways to build a student predictive model through integrating linked data sources that are internal or external to higher education institutions, without having to rely on questionnaire data that have been essential in existing models. The traditional questionnaire data, which are not available in the institutional internal databases, they will be replaced by any credible data from external data sources as many as possible.

**2.6 (M*) Key research question** (*Specify hypothesis if applicable*):

Key research question:

1. Are linked data technologies well suited to address student retention, progression and completion? What are the advantages using linked data technologies in this respect? Can we show how student retention, progression and completion can be efficiently addressed by sharing information using linked data technologies?

Hypotheses:

a) it is possible to provide accurate student prediction models by combining internal and external data sources. b)Linked data is efficient to support building student prediction model when combining internal/external data sources, and c)Internal/external data sources can be used to compensate the lack of questionnaire data in building student prediction model.

**2.7 (M*) Study design** (*Give a **brief** outline of basic study design*)

*Outline what approach is being used, why certain methods have been chosen.*

An experimental platform has been developed which is able to convert raw data to RDF and also able to integrate data from disparate data sources to make a final set of data to build the predictive model. Logistic regression will be used to develop the predictive model. To evaluate the experimental platform to build the predictive model and to test the model accuracy, we need individual student data. As per the current rules and regulations of the University of Southampton, we are not eligible to use university databases to collect the required individual student information that's why we need to conduct questionnaire to collect the individual student information. We are intending to collect around 100-150 students data so that 70% of the students data will be used to build the predictive model and the rest of the students (30%) data will be used to validate the model accuracy. We will not require to communicate further with the participants of this study once they complete the questionnaire.

## 3. Sample and Setting

**3.1 (M*) How are participants to be *approached*?** *Give details of what you will do if recruitment is insufficient. If participants will be accessed through a third party (e.g. children accessed via a school) state if you have permission to contact them and **upload any letters of agreement to your submission in ERGO**.*

I am expecting at least 100 to 150 1st degree/undergraduate students who have been enrolled in the academic year 2010/2011 in any programmes of study in the university of Southampton to participate in my study. The invitation with my contact details, online survey URL and participants information will be sent through email to the prospective participants. Also advertisements will be done by circulating posters in all around the university especially around heartly library, SUSU, interchange, Nuffield theatre area. If there is a lower participaion rate than I expected the time will be extended to collect more data until the student number I expected.

**3.2 (M*) Who are the proposed sample and where are they from (e.g. fellow students, club members)?** *List inclusion/exclusion criteria if applicable. NB The University does not condone the use of 'blanket emails' for contacting potential participants (i.e. fellow staff and/or students).*

*It is usually advised to ensure groups of students/staff have given prior permission to be contacted in this way, or to use of a third party to pass on these requests. This is because there is a potential to take advantage of the access to 'group emails' and the relationship with colleagues and subordinates; we therefore generally do not support this method of approach.*

*If this is the only way to access a chosen cohort, a reasonable compromise is to obtain explicit approval from the Faculty Ethics Committee (FEC) and also from a senior member of the Faculty in case of complaint.*

Participants are the 1st degree/undergraduate students in the university of Southampton who enrolled in 2010/2011 academic year in any programme of study.

**3.3 (M*) Describe the relationship between researcher and sample** (*Describe any relationship e.g. teacher, friend, boss, clinician, etc.*)

Participants are the fellow students.

**3.4 (M*) Describe how you will ensure that fully informed consent is being given:** (*include how long participants have to decide whether to take part*)

I will provide detail participant information at the first instance in the invitation email so that participants can read and know about the study. Also I will give my contact information in the email so that participants can contact me if they have any more queries about the study. In the email I will request them to read the information about the study first then if they are happy to take part they will require to follow the provided questionnaire URL to start the questionnaire session. Before starting the questionnaire they require to tick a check box to inform their consent. Also the participants who will see the advertisement through poster and wish to participate in this study they require to follow the provided URL in the poster. In the URL, they will find the participation information to read at first and if they are happy to take part they need to tick the box to inform their consent before they start the questionnaire.

## 4. Research Procedures, Interventions and Measurements

**4.1 (M*) Give a brief account of the procedure as experienced by the participant**

*(Make clear who does what, how many times and in what order. Make clear the role of all assistants and collaborators. Make clear total demands made on participants, including time and travel). **Upload any copies of questionnaires and interview schedules to your submission in ERGO**.*

The participants who wiling to participate in this study need to give their consent by ticking the check box before starting the questionnaire session in online to inform their consent. Then they need to complete the questionnaire following the provided URL, there are 6 sections in the questionnaire which requires around 20-30 minutes to complete. I am expecting around 100-150 participants. There will be a draw and 50 of the participants who will complete the questionnaire will have a 20 pound amazon voucher. The participants are required to provide their email address in a hidden questionnaire which separate from the main questionnair to participate in the draw and to contact the winner to give their voucher. However, once they have completed the study, I will delete this information from the datafile so that all of their responses are anonymous. No more contact will be made with them once they complete the study.

The questionnaire session will follow the following steps:

Step 1: The participant will be provided a study information to read in the invitation email.

Step 2: After reading the study information in the email if they are willing to participate, they need to follow the questionnaire URL to take part in the questionnaire session.

Step 3: The participation information can also be found at the top of the questionnaire to read. After reading the information participants need to inform their cosent by ticking the check box to start their questionnaire session.

Step 4: There are 6 sections in the questionnaire which will take around 35-40 minutes to complete.

Step 5: Affter completing section 6 the participants need to follow a link (hidden questionnaire URL) to provide their email address to take part in the draw to win the 20-pound amazon voucher.

During the study time if the participants don't want to continue they are welcome and can leave the study any time they wish. The URL of the questionnaire is https://www.isurvey.soton.ac.uk/5072.

## 5. Study Management

**5.1 (M*) State any potential for psychological or physical discomfort and/or distress?**

There is no psychological or physical discomfort and distress assosiated in thus study.

**5.2 (M*) Explain how you intend to alleviate any psychological or physical discomfort and/or distress that may arise? (if applicable)**

N/A

**5.3 Explain how you will care for any participants in 'special groups'** *(i.e. those in a dependent relationship, vulnerable or lacking in mental capacity)* **(if applicable)?**

N/A

**5.4 Please give details of any payments or incentives being used to recruit participants (if applicable)?**

Incentives will be given to all participants who successfully complete the questionnaire. Each participant will get a 20-pound Amazon Voucher.

**5.5 i) How will participant anonymity and/or data anonymity be maintained (if applicable)?**

*Two definitions of anonymity exist:*

*i)* Unlinked anonymity - *Complete anonymity can only be promised if questionnaires or other requests for information are not targeted to, or received from, individuals using their name or address or any other identifiable characteristics. For example if questionnaires are sent out with no possible identifiers when returned, or if they are* *picked up by respondents in a public place, then anonymity can be claimed. Research methods using interviews cannot usually claim anonymity – unless using telephone interviews when participants dial in.*

*ii)* Linked anonymity - *Using this method, complete anonymity cannot be promised because participants can be identified; their data may be coded so that participants are not identified by researchers, but the information provided to participants should indicate that they could be linked to their data.*

Participants will not be required to provide their student id, name or full address. I am expecting around 100-150 participants and the data will be labelled with a auto generated id number for each participants. All information will be kept securely in my password protected personal computer in the university of Southampton. I will not link the participants information to any other information so that they can be identified. The participants' email address will not be shared with participants or any third party. No further communication will be made with the participants once the questionnaire session has been done.

**5.5 ii) How will participant confidentiality be maintained (if applicable)?**

*Confidentiality is defined as the non-disclosure of research information except to another authorised person. Confidential information can be shared with those who are already party to it, and may also be disclosed where the person providing the information provides explicit consent.*

In the session, an online questionnaire will be distributed amongst the participants to collect quantitative data to develop the predictive model. It does not require participants to reveal any information such as names, student id, full address by which they can be personally identified. Though the participants email address will be collected for the draw using a hidden questionnaire so that there will not exist any link between the collected data and the email address. Though participants are required to provide their gender, ethnicity etc information in the questionnaire. However they are required to inform their consent before starting the questionnaire. The raw data of the research will not be made available to the participants or any third party. The data will be securely maintained in a password protected parsonal computer in the university of Southampton.

**5.6 (M*) How will personal data and study results be stored securely during and after the study?** *Researchers should be aware of, and compliant with, the Data Protection policy of the University. You must be able to demonstrate this in respect of handling, storage and retention of data.*

In the session, an online questionnaire will be distributed amongst the participants to collect quantitative data to develop the predictive model. It does not require participants to reveal any information such as names, student id, full address by which they can be personally identified. Though the participants email address will be collected for the draw with another hidden questionnaire so that there will not exist any link between the collected data and the email address. Though participants are required to provide their gender, ethnicity etc information in the questionnaire. However they are required to inform their consent before starting the questionnaire. The raw data of the research will not be made available to the participants or any third party. The data will be securely maintained in a password protected parsonal computer in the university of Southampton.

**5.7 (M*) Who will have access to these data?**

No one can access the data except me.

# Appendix C

## List of Variables Related to Student Retention, Progression and Completion

Based on the literature, a list of variables related to student retention, progression and completion is presented in the following table, which are categorized into five terminologies, such as students' individual attribute, academic preparedness, academic variable, support and institutional variable.

Table C.1 List of variables related to student retention, progression and completion

| List of Variables |
|---|
| **Students' Individual attribute** |
| Age |
| Gender |
| Ethnicity |
| Disability |
| Residence/ Domicile |
| Accommodation |
| Interests |
| Language preference |
| Life circumstances |
| -Dependents/Family concern |
| -Childcare |
| Work circumstances (current) |
| -Part time/Full time |
| -Inside the campus/ outside the campus |
| - Paid / unpaid work |
| Work experience prior to entering HE |
| 1st in family to enter HE |
| Reason to enter the HE |
| Goals and commitment |
| -Qualification sought |
| -Career goal |
| Financial situation |
| -Financing study |

- -Student loan (if took loan when before entering in HE or after)
- -Fee paid/not paid
- Family background
  - -Family income/Socio-economic status
  - -Parental educational level
  - -Parental expectation

**Academic preparedness**

- High School GPA/ Entry qualification
- Prior knowledge of HEI/ programme
  - -Academic rules
  - -Social rules
  - -Course requirements
  - -Curriculum
  - -Graduation requirements
- Making the right choice
  - -Institution
  - -Programme
- Realistic expectations
- Motivation
- Special test result
  - -Aptitude test
- Entry route

**Academic variable**

- Programme of study
- Year of study
- Subjects studied in the first year
- Flexibility
  - -Timetable
  - -Deadline
  - -Opportunities for re-taking courses
- Class size
- Credit load/Workload of course
- Teaching and Learning style
  - -Active learning
  - -Student centred
  - -Practical project
  - -Problem based learning
- Feedback
- Assessment
  - -Coursework assessment
  - -Peer assessment
  - -On-line assessment
- Student attendance

|  |
|---|
| Record of academic achievements<br>-Assessment scores<br>- Transcripts (grades)<br>Student satisfaction<br>Activities/Skills/experience/competency/knowledge<br>-Involvement in campus programs (freshmen orientation course, educational opportunities program)<br>-Group activities<br>Type of students/Study mode<br>-Full Time/ Part Time<br>First destination |
| **Support**<br>-Student-student interaction / Peer support<br>-Contact time with staff / Tutor support<br>-Support from friends and family<br>-Need to reject past attitude and value |
| **Institutional variable**<br>Size and type of institutions<br>Location of the university<br>-Travelling<br>Institutional expenditure/ Budgeting and funding<br>-Instruction (faculty, teaching)<br>-Academic support (libraries, academic computing)<br>-Student service (admissions, register, student development offices)<br>-Institutional support (administrative, legal, executive, expenditures)<br>-Institutional grants (merit and need based scholarship) |

# Appendix D

## SRPC Ontology

The following table presents all the classes and properties of SRPC ontology.

Table D.1 Classes and properties of SRPC ontology

| Term Name | Type | Definition |
|---|---|---|
| Student | class | A student in a university. |
| Programme | class | A programme provided by the university. |
| Major | class | A major in a programme provided by the university. |
| Module | class | A module provided by the university. |
| programID | property | The identification of the programme. |
| majorID | property | The identification of the major. |
| programName | property | The name of the programme. |
| majorName | property | The name of the Major. |
| moduleID | property | The identification of the module. |
| moduleName | property | The name of the module. |
| inProgram | property | A programme in which a student enrolled in. |
| takingMajor | property | A major that a student studied in. |
| takingModule | property | A module that a student studied. |
| studyMode | property | The study mode of a student. |
| inYear | property | The current year of study of the student (First year/Second year) Third year/Fourth year). |
| parentsHEQual | property | The status of whether a student's parents have HE or not. |
| attndInduction | property | The status of whether a student attended induction programme or not. |
| stdNeighborhood | property | The neighborhood of a student from which this student has come. |
| typeOfStudent | property | The type of student whether he/she is young or mature. |
| sexOfStudent | property | The sex of student whether he/she is male or female. |
| maritalStatus | property | The marital status of a student. |
| admissionType | property | The admission type of a student. |
| hasDisability | property | Whether a student has any disability or not. |
| hasIndicator | property | Whether a student is at-risk or not at-risk in his/her |
| motherOccupation | property | The occupation of a student's mother |
| fatherOccupation | property | The occupation of a student's father |

# Appendix E

## Details of the Variables for Experiment 1

The following is an enhance list of all variables for Table 7.1 with variable definitions which has been used in experiment 1.

Table E.1 List of variables, variables definition and variable sources

| Variables | Variable Definition | Variable source |
|---|---|---|
| **Gender** | Male =1, Female = 2 | IDB |
| **Ethnicity** | White=1, Non-White=0 | IDB |
| **A Level tariff points** | A*=140, A=120, B= 100, C=80, D=60<br>Example, if a student's A level grades are AAA then his A level tariff points counted as AAA=120+120+120=360 | IDB |
| **Accommodation Type** | University halls=1, Others=2 | IDB |
| **First year's first semester marks** | 71-100=1, 61-70=2, 51-60=3, 41-50=4 | IDB |
| **Source of tuition fee** | Grant/Scholarship = 1, Student loan = 2, Family/Yourself = 3 | IDB |
| **Study field** | Applied (engineering, physics, chemistry etc) =1, Non-applied (Languages etc) = 0 | IDB |
| **Parents' have HE qualification** | Yes=1, No=0 | IDB |
| **Peer Group interaction (7 items/variables)**<br>1. Since coming to this university, I have made close personal relationship with other students.<br>2. The student friendships I have developed at the university have been personally satisfying.<br>3. My interpersonal relationships with other students have had a positive influence on my personal growth, attitudes, and values.<br>4. My interpersonal relationships with other students have had a positive influence on my intellectual growth and interest in ideas.<br>**5.** It has been difficult for me to meet and make friends with other students.<br>**6.** Few of the students I know would be willing to listen to me and help me if I had a personal problem.<br>**7.** Most students at this university have values and attitudes different from my own. | Ordinal variables, Ranking from 1 to 4. Strongly agree=4, Agree=3, Disagree=2 and Strongly disagree=1 | Questionnaire (IIS) |

*Appendix E*

| Variables | Variable Definition | Variable source |
|---|---|---|
| **Student-Faculty interaction (5 items/variables)**<br>1. My non-classroom interactions with faculty have had a positive influence on my personal growth, values and attitudes.<br>2. My non-classroom interactions with faculty have had a positive influence on my intellectual growth and interest in ideas.<br>3. My non-classroom interactions with faculty have had a positive influence on my career goals and aspirations.<br>4. Since coming to this university, I have developed a close, personal relationship with at least one faculty member.<br>5. I am satisfied with the opportunities to meet and interact informally with faculty members. | Ordinal variable, Ranking from 1 to 4. Strongly agree=4, Agree=3, Disagree=2 and Strongly disagree=1 | Questionnaire (IIS) |
| **Faculty Concern For Student Development and Teaching (5 items/variables)**<br>1. Few of the faculty members I have had contact with are generally interested in students.<br>2. Few of the faculty members I have had contact with are generally outstanding and superior teachers.<br>3. Few of the faculty members I have had contact with are willing to spend time outside of class to discuss issues of interest and importance to students.<br>4. Most of the faculty members I have had contact with are interested in helping students grow in more than just academic areas.<br>5. Most faculty members I have had contact with are genuinely interested in teaching. | Ordinal variable, Ranking from 1 to 4. Strongly agree=4, Agree=3, Disagree=2 and Strongly disagree=1 | Questionnaire (IIS) |
| **Academic and Intellectual Development (7 items/variables)**<br>1. I am satisfied with the extent of my intellectual development since enrolling in this university.<br>2. My academic experience has had a positive influence on my intellectual growth and interest in ideas.<br>3. I am satisfied with my academic experience at this university.<br>4. Few of my courses this year have been intellectually stimulating.<br>5. My interest in ideas and intellectual matters has increased since coming to this university<br>6. I am more likely to attend a cultural event now than I was before coming to this university.<br>7. My academic performance has met my expectation. | Ordinal variable, Ranking from 1 to 4. Strongly agree=4, Agree=3, Disagree=2 and Strongly disagree=1 | Questionnaire (IIS) |
| **Institutional Commitment I**<br>- Your choice of this institution was? | 1$^{st}$ =1, 2$^{nd}$=2,Others=3 | Questionnaire (IIS) |
| **Goal Commitment I**<br>- What is the highest expected academic degree? | Master's degree or above=2, Bachelor's degree or below=1 | Questionnaire (IIS) |
| **Institutional Commitment II (2 items/variables)**<br>1. It is important for me to graduate from this university.<br>2. I am confident that I have made the right decision in choosing to attend this university. | Ordinal variable, Ranking from 1 to 4. Strongly agree=4, Agree=3, Disagree=2 and Strongly disagree=1 | Questionnaire (IIS) |

| Variables | Variable Definition | Variable source |
|---|---|---|
| **Goal Commitment II**<br>• Getting good result is not important to me | Ordinal variable, Ranking from 1 to 4. Strongly agree=4, Agree=3, Disagree=2 and Strongly disagree=1 | Questionnaire (IIS) |
| **Intention**<br>• It is likely that I will register at this university next year. | Ordinal variable, Ranking from 1 to 4. Strongly agree=4, Agree=3, Disagree=2 and Strongly disagree=1 | Questionnaire (IIS) |
| **The teaching on my course (4 items/variables)**<br>1. Staff are good at explaining things.<br>2. Staff have made the subject interesting.<br>3. Staff are enthusiastic about what they are teaching.<br>4. The course is intellectually stimulating. | Numeric (%) | EDB (Unistats) |
| **Assessment and feedback (5 items/variables)**<br>1. The criteria used in marking have been clear in advance.<br>2. Assessment arrangements and marking have been fair.<br>3. Feedback on my work has been prompt.<br>4. I have received detailed comments on my work.<br>5. Feedback on my work has helped me clarify things I did not understand. | Numeric (%) | EDB (Unistats) |
| **Academic support (3 items/variables)**<br>1. I have received sufficient advice and support with my studies.<br>2. I have been able to contact staff when I needed to.<br>3. Good advice was available when I needed to make study choices. | Numeric (%) | EDB (Unistats) |
| **Personal development (3 items/variables)**<br>1. The course has helped me present myself with confidence.<br>2. My communication skills have improved.<br>3. As a result of the course, I feel confident in tackling unfamiliar problems. | Numeric (%) | EDB (Unistats) |
| **Overall, I am satisfied with the quality of the course.** | Numeric (%) | EDB (Unistats) |

# Appendix F

## Derivation of students' parents' annual mean income and SEC based on SOC 2010

The following is the full derivation table for students' parents' annual mean income and students' socio economic class (SEC) by linking students both parents' occupation with office for national statistics (ONS) published annual gross income and SEC open datasets based on Standard Occupational Classification 2010 (SOC 2010).

Table F.1 Derivation of students' parents' annual income and SEC based on SOC 2010.

| Mother's occupation | ONS occupation | Mother's annual Mean Gross income | Father's occupation | ONS occupation | Father's annual Gross income | SEC |
|---|---|---|---|---|---|---|
| Deputy Headteacher | Education advisers and school inspectors | 33,698 | Technical Director | Information technology and telecommunications directors | 64,141 | 2 |
| Teacher | Primary and nursery education teaching professionals | 29,657 | Retired | Not Classified | | 2 |
| Teacher | Primary and nursery education teaching professionals | 29,657 | Senior Manager | Office managers | 41,549 | 2 |
| Sales | Sales and retail assistants | 8,649 | Director | Functional managers and directors n.e.c. | 62,726 | 2 |
| Clerk | Records clerks and assistants | 16,873 | Electrician | Electricians and electrical fitters | 29,741 | 5 |

| Mother's occupation | ONS occupation | Mother's annual Mean Gross income | Father's occupation | ONS occupation | Father's annual Gross income | SEC |
|---|---|---|---|---|---|---|
| Business Services | Business and related associate professionals n.e.c. | 23,212 | NHS Provisioning Mgr | Health services and public health managers and directors | 53,867 | 1.1 |
| Photographer | Photographers, audio-visual and broadcasting equipment operators | 19,074 | Photographer | Photographers, audio-visual and broadcasting equipment operators | 26,619 | 4 |
| Lecturer | Higher education teaching professionals | 32,600 | Company Director | Functional managers and directors n.e.c. | 62,726 | 2 |
| Public service | Public services associate professionals | 25,735 | Bank employee | Bank and post office clerks | 25,893 | 3 |
| Teaching Assistant | Teaching assistants | 11,457 | C&I Engineer | Engineering professionals n.e.c. | 41,297 | 1.2 |
| Teaching Assistant | Teaching assistants | 11,457 | Civil Engineer | Civil engineers | 39,341 | 1.2 |
| Unknown | Not Classified | | doctor | Medical practitioners | 83,760 | 1.2 |
| Unknown | Not Classified | | Investment Banker | Finance and investment analysts and advisers | 64,042 | 2 |
| Accountant | Chartered and certified accountants | 31,005 | Sales manager | Sales accounts and business development managers | 55,601 | 1.2 |
| Carer p-t retail | Retail cashiers and check-out operators | 8,276 | Firefighter | Fire service officers (watch manager and below) | 26,087 | 3 |
| Secretary | Company secretaries | 16,345 | Coach Operator | Bus and coach drivers | 22,437 | 7 |
| Biochemist | Biological scientists and biochemists | 31,925 | Architect | Architects | 44,377 | 1.2 |

| Mother's occupation | ONS occupation | Mother's annual Mean Gross income | Father's occupation | ONS occupation | Father's annual Gross income | SEC |
|---|---|---|---|---|---|---|
| Housewive | Not Classified | | Businessman | Business and related associate professionals n.e.c. | 37,508 | 2 |
| Self Employed Consultant | Management consultants and business analysts | 37,349 | IT Architect | IT business analysts, architects and systems designers | 42,686 | 1.2 |
| Accountant | Chartered and certified accountants | 31,005 | Financial Adviser | Finance and investment analysts and advisers | 64,042 | 2 |
| Nurse | Nurses | 25,474 | Social Care worker | Social workers | 28,778 | 2 |
| Artist | Artists | 22,427 | Control Manager | Production managers and directors in manufacturing | 55,625 | 1.1 |
| None | Not Classified | | Construction worker | Construction operatives n.e.c. | 21,262 | 7 |
| Housewife | Not Classified | | Taxi Driver | Taxi and cab drivers and chauffeurs | 16,572 | 4 |
| Administration | Other administrative occupations n.e.c. | 14,870 | Software Developer | Programmers and software development professionals | 38,947 | 1.2 |
| Cashier | Retail cashiers and check-out operators | 8,276 | Assistant Manager | Office managers | 41,549 | 2 |
| Nurse | Nurses | 25,474 | Chemical Engineer | Chemical scientists | 34,939 | 1.2 |
| Midday supervisor | Sales supervisors | 13,571 | insolvency examiner | Business and related associate professionals n.e.c. | 37,508 | 2 |
| Bank manager | Financial accounts managers | 31,739 | Manager | Office managers | 41,549 | 2 |

*Appendix F*

| Mother's occupation | ONS occupation | Mother's annual Mean Gross income | Father's occupation | ONS occupation | Father's annual Gross income | SEC |
|---|---|---|---|---|---|---|
| Osteopath | Medical and dental technicians | 22,129 | research consultant | Research and development managers | 49,400 | 1.2 |
| Teacher | Primary and nursery education teaching professionals | 29,657 | Salesman | Sales and retail assistants | 12,831 | 2 |
| Maths Teacher | Primary and nursery education teaching professionals | 29,657 | IT Consultant | IT business analysts, architects and systems designers | 42,686 | 1.2 |
| Specialist Nurse | Nurses | 25,474 | Legal Services | Legal professionals n.e.c. | 76,835 | 1.2 |
| FE lecturer | Further education teaching professionals | 25,719 | Software engineer | Engineering professionals n.e.c. | 41,297 | 1.2 |
| Carer | Care workers and home carers | 12,545 | Disabled/Unemployed | Not Classified | | 6 |
| School Assistant | Educational support assistants | 11,351 | Programmer | Programmers and software development professionals | 38,947 | 1.2 |
| Retired | Not Classified | | Manager | Office managers | 41,549 | 2 |
| House wife | Not Classified | | Machanical engineer | Mechanical engineers | 43,715 | 1.2 |
| Admin clerk | Records clerks and assistants | 16,873 | Sergeant in the army | Police officers (sergeant and below) | 40,563 | 3 |
| Childcare Worker | Childcare and related personal services | 11,314 | Carpenter | Carpenters and joiners | 22,904 | 4 |
| Teacher | Primary and nursery education teaching professionals | 29,657 | Businessman | Business and related associate professionals n.e.c. | 37,508 | 2 |
| Book keeper | Book-keepers, payroll managers and wages clerks | 18,118 | Factory Worker | Sheet metal workers | 25,251 | 6 |

| **Mother's occupation** | **ONS occupation** | **Mother's annual Mean Gross income** | **Father's occupation** | **ONS occupation** | **Father's annual Gross income** | **SEC** |
|---|---|---|---|---|---|---|
| Painter | Painters and decorators | | Senior Manager | Office managers | 41,549 | 2 |
| HLTA | Special needs education teaching professionals | 27,834 | Driving Instructor | Driving instructors | 24,757 | 2 |
| Educational officer | Education advisers and school inspectors | 33,698 | Educational Officer | Education advisers and school inspectors | 44,712 | 1.2 |
| House Wife | Not Classified | | Electronic Engineer | Electronics engineers | 38,799 | 1.2 |
| Secretary | Company secretaries | 16,345 | Company Director | Elementary trades and related occupations | 19,208 | 6 |
| Catering | Restaurant and catering establishment managers and proprietors | 20,636 | Pollution Officer | Health and safety officers | 35,457 | 2 |
| Teacher | Primary and nursery education teaching professionals | 29,657 | Unknown | Not Classified | | 2 |
| Parliamentary Casewo | Social workers | 26,905 | Engineer | Engineering professionals n.e.c. | 41,297 | 1.2 |
| Teaching Assistant | Teaching assistants | 11,457 | Train Driver | Train and tram drivers | 41,377 | 5 |
| N/A | Not Classified | | Financial Advisor | Finance and investment analysts and advisers | 64,042 | 2 |
| None | Not Classified | | Management | Management consultants and business analysts | 51,990 | 1.2 |
| Teacher | Primary and nursery education teaching professionals | 29,657 | Technical Director | Information technology and telecommunica tions directors | 64,141 | 2 |

*Appendix F*

| Mother's occupation | ONS occupation | Mother's annual Mean Gross income | Father's occupation | ONS occupation | Father's annual Gross income | SEC |
|---|---|---|---|---|---|---|
| Sales Assistant | Sales and retail assistants | 8,649 | Programmer | Programmers and software development professionals | 38,947 | 1.2 |
| Financial Director | Financial institution managers and directors | 40,567 | Director | Functional managers and directors n.e.c. | 62,726 | 2 |
| Volunteer | Youth and community workers | 19,268 | Sales Engineer | Sales related occupations n.e.c. | 27,873 | 3 |
| House wife | Not Classified | | Radiographer | Medical radiographers | 37,917 | 2 |
| Librarian | Librarians | 21,869 | Auditer | Quality assurance technicians | 27,660 | 5 |
| Receptionist | Receptionists | 11,953 | Unknown | Not Classified | | 6 |
| Teaching Assistant | Teaching assistants | 11,457 | OSS Architect | Architects | 44,377 | 1.2 |
| LSA | Educational support assistants | 11,351 | IT | IT engineers | 29,530 | 3 |
| Supply Teacher | Teaching and other educational professionals n.e.c. | 17,447 | Chartered Accountant | Chartered and certified accountants | 44,240 | 1.2 |
| Speech Therapist | Speech and language therapists | 25,047 | University Librarian | Librarians | 25,180 | 2 |
| Shop Assistant | Sales and retail assistants | 8,649 | Payroll manager | Book-keepers, payroll managers and wages clerks | 28,955 | 3 |
| Corporate Affairs | Officers of non-governmental organisations | 19,352 | Property Developer | Property, housing and estate managers | 45,374 | 1.1 |
| Nurse | Nurses | 25,474 | Technician | Other elementary services occupations n.e.c. | 11,183 | 2 |

| Mother's occupation | ONS occupation | Mother's annual Mean Gross income | Father's occupation | ONS occupation | Father's annual Gross income | SEC |
|---|---|---|---|---|---|---|
| Occupational Therapist | Occupational therapists | 25,769 | N/A | Not Classified | | 2 |
| Unemployed | Not Classified | | Patent Specialist | | | |
| Supermarket cashier | Retail cashiers and check-out operators | 8,276 | Hands out surveys | Not Classified | | 6 |
| N/A | Not Classified | | Area Manager | Marketing and sales directors | 90,132 | 1.1 |
| Accountant | Chartered and certified accountants | 31,005 | Accountant | Chartered and certified accountants | 44,240 | 1.2 |
| Teaching Assistant | Teaching assistants | 11,457 | Community Nurse | Nurses | 29,642 | 2 |
| Teaching Advisor | Teaching assistants | 11,457 | Project Manager | IT project and programme managers | 52,766 | 1.2 |
| Nurse | Nurses | 25,474 | Businessman | Business and related associate professionals n.e.c. | 37,508 | 2 |
| Manager | Office managers | 25,771 | Manager | Office managers | 41,549 | 2 |
| None | Not Classified | | Banker | Bank and post office clerks | 25,893 | 3 |
| IT technician | IT operations technicians | 24,647 | retired | Not Classified | | 2 |
| Debt Advisor | Debt, rent and other cash collectors | 14,907 | Chief Executive | Chief executives and senior officials | 140,330 | 1.1 |
| Library Assistant | Library clerks and assistants | 11,816 | Genealogist | Physical scientists | 48,505 | 1.2 |
| Teaching Assistant | Teaching assistants | 11,457 | Community Nurse | Nurses | 29,642 | 2 |
| Architect | Architects, town planners and surveyors(2431= architect) | 27,899 | Consulting Engineer | Engineering professionals n.e.c. | 41,297 | 1.2 |
| Accounts assistant | Financial and accounting technicians | 32,078 | Entrepeneur | Business and financial project management professionals | 53,651 | 1.2 |

*Appendix F*

| Mother's occupation | ONS occupation | Mother's annual Mean Gross income | Father's occupation | ONS occupation | Father's annual Gross income | SEC |
|---|---|---|---|---|---|---|
| Shop cashier | Retail cashiers and check-out operators | 8,276 | Hands out surveys | Not Classified | | 6 |
| Teaching Assistant | Teaching assistants | 11,457 | Pilot | Aircraft pilots and flight engineers | 74,209 | 1.2 |
| Teacher | Primary and nursery education teaching professionals | 29,657 | Physicist | Physical scientists | 48,505 | 1.2 |
| Data Manager | Managers and directors in storage and warehousing | 25,964 | Head of developemnt | Research and development managers | 49,400 | 1.2 |
| Cleaner | Cleaners and domestics | 6,938 | Chef | Chefs | 17,845 | 5 |
| Nurse | Nurses | 25,474 | Chemical Engineer | Chemical scientists | 34,939 | 1.2 |
| Business partner | Sales accounts and business development managers | 37,826 | IFA | Finance and investment analysts and advisers | 64,042 | 2 |
| Communicatio ns | Communication operators | 21,789 | Business Analyst | Management consultants and business analysts | 51,990 | 1.2 |
| Primary Teaching | Primary and nursery education teaching professionals | 29,657 | Accountant | Chartered and certified accountants | 44,240 | 1.2 |
| Teacher | Primary and nursery education teaching professionals | 29,657 | Accountant | Chartered and certified accountants | 44,240 | 1.2 |
| Maths Teacher | Primary and nursery education teaching professionals | 29,657 | Doctor | Medical practitioners | 83,760 | 1.2 |
| Artist | Artists | 22,427 | Tube factory | Sheet metal workers | 25,251 | 6 |

| Mother's occupation | ONS occupation | Mother's annual Mean Gross income | Father's occupation | ONS occupation | Father's annual Gross income | SEC |
|---|---|---|---|---|---|---|
| House Wife | Not Classified | | Self Employed | Not Classified | | |
| Administration | Other administrative occupations n.e.c. | 14,870 | Bank Manager | Financial accounts managers | 64,149 | 2 |
| Solicitors | Solicitor | 41,243 | Solicitor | Solicitors | 57,177 | 1.2 |
| Teacher | Primary and nursery education teaching professionals | 29,657 | Performance Manager | Functional managers and directors n.e.c. | 62,726 | 2 |
| Teaching Assistant | Teaching assistants | 11,457 | Financial Advisor | Finance and investment analysts and advisers | 64,042 | 2 |
| Teacher | Primary and nursery education teaching professionals | 29,657 | Civil engineer | Civil engineers | 39,341 | 1.2 |
| Data Manager | Managers and directors in storage and warehousing | 25,964 | systems artitect | IT business analysts, architects and systems designers | 42,686 | 1.2 |
| Teacher | Higher education teaching professionals | 32,600 | Engineer | Engineering professionals n.e.c. | 41,297 | 1.2 |
| Consultant Anesthetist | Nurses | 25,474 | Pharmaceutical Representative | Sales and retail assistants | 8,649 | 2 |
| Cleaner | Cleaners and domestics | 6,938 | Youth worker | Youth and community workers | 20,865 | 2 |
| Procurement | Buyers and procurement officers | 27,954 | Estate Agent | Estate agents and auctioneers | 29,328 | 2 |
| Meteorologist | Natural and social science professionals n.e.c. | 32,964 | Meteorologist | Natural and social science professionals n.e.c. | 38,134 | 1.2 |

*Appendix F*

| Mother's occupation | ONS occupation | Mother's annual Mean Gross income | Father's occupation | ONS occupation | Father's annual Gross income | SEC |
|---|---|---|---|---|---|---|
| Finance Officer | Finance officers | 21,050 | Electrician | Electricians and electrical fitters | 29,741 | 5 |
| N/A | Not Classified | | Architect | Architects | 44,377 | 1.2 |
| Retail | Sales and retail assistants | 8,649 | Labourer | Other elementary services occupations n.e.c. | 11,183 | 7 |
| N/A | Not Classified | | Auto parts courier | Postal workers, mail sorters, messengers and couriers | 21,104 | 6 |
| Clerk | Records clerks and assistants | 16,873 | Engineer | Engineering professionals n.e.c. | 41,297 | 1.2 |
| Stone conservator | Conservation professionals | 23,891 | consultant (google) | Management consultants and business analysts | 51,990 | 1.2 |
| Asst. Photographer | Photographers, audio-visual and broadcasting equipment operators | 19,074 | Photographer | Photographers, audio-visual and broadcasting equipment operators | 26,619 | 4 |
| N/A | Not Classified | | Doctor | Medical practitioners | 83,760 | 1.2 |
| School Nurse | Nurses | 25,474 | IT Manager | IT specialist managers | 38,598 | 1.2 |
| N/A | Not Classified | | Engineer | Engineering professionals n.e.c. | 41,297 | 1.2 |
| Technician | Other elementary services occupations n.e.c. | 8,998 | Minister of Religion | Clergy | 21,764 | 1.2 |
| Book keeper | Book-keepers, payroll managers and wages clerks | 18,118 | car insurance | Pensions and insurance clerks and assistants | 30,564 | 3 |
| Admin clerk | Records clerks and assistants | 16,873 | military police | Officers in armed forces | 53,778 | 1.1 |

| Mother's occupation | ONS occupation | Mother's annual Mean Gross income | Father's occupation | ONS occupation | Father's annual Gross income | SEC |
|---|---|---|---|---|---|---|
| Marketing company | Marketing associate professionals | 25,065 | credit insurance | Insurance underwriters | 47,986 | 1.2 |
| Social Worker | Social workers | 26,905 | Small Business Owner | Elementary trades and related occupations | 19,208 | 2 |
| Security officer | Security guards and related occupations | 21,360 | Managing director | Elementary trades and related occupations | 19,208 | 6 |
| Waitress | Waiters and waitresses | 6,555 | Software developer | Programmers and software development professionals | 38,947 | 1.2 |
| Home Carer | Care workers and home carers | 12,545 | Director | Functional managers and directors n.e.c. | 62,726 | 2 |