UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

KNOWLEDGE BASED STUDENT ACADEMIC ADVISING IN INSTITUTIONS OF HIGHER LEARNING IN KENYA

BY

MOSES KARANI

P58/63367/2011

SUPERVISOR

CHRISTOPHER MOTURI

SEPTEMBER 2013

Submitted in partial fulfillment of the requirements of the Master of Science in Computer Science

DECLARATION

I hereby declare that the research project presented in this report is my original work and has not been presented in any other institution. Due and clear reference is made to the works of other researches that have informed this project.

Signature_____

Date_____

Karani Moses Orupia

P58/63367/2011

The Research Project has been submitted in partial fulfillment of the Requirements of the Degree of Master of Science in Computer Science at the University of Nairobi with my approval as the University Supervisor.

Signature_____

Date_____

Mr. Christopher Moturi

School of Computing and Informatics

University of Nairobi

DEDICATION

To

All my teachers who have taught and counseled me over the years,

My father and late mother,

Daisy and Valerie

Above all to God

ACKNOWLEDGEMENT

The counsel and advice of my supervisor Mr. Christopher Moturi is highly appreciated. I also appreciate the useful advice I got from my evaluation panel members and fellow classmates as a whole. Above all to God be the glory.

ABSTRACT

Technological advances in the last two decades have led to reduced costs of computer hardware and software. This has in turn led to widespread acquisition and use Student Management Information Systems by institutions of higher learning resulting in huge databases of student data. Despite these advancements, reporting applications have remained largely rudimentary employing simple reports such as a student transcript, a consolidated mark sheet and so on.

Data mining techniques can be used to obtain knowledge from large collections of data by employing techniques from both computer science and statistics thereby enriching the reporting applications. These techniques can be used to provide patterns and analyses that cannot be obtained using rudimentary querying and reporting techniques. Furthermore these patterns can be used to build models that can be used to perform predictions among many other applications.

This research proposes to employ data mining techniques to build predictive models that can be used to predict the degree honors class that a student is likely to get. Using this knowledge plus that obtained using clustering and classification data mining techniques, prudent advice can be furnished to students early enough about their likely final degree honors class. With this knowledge, students can then be advised on how to improve their performance using the knowledge provided by classification and clustering algorithms.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

CRISP-DM – Cross Industry Standard Process for Data Mining

CRM – Customer Relationship Management

EDM – Educational Data Mining

KDD – Knowledge Discovery in Databases

OLAP – Online Analytical Processing

SMIS – Student Management Information System

SEMMA – Sample Explore Modify Model Asses

CHAPTER 1

INTRODUCTION

1.0 Background

Prior to the introduction of computer based student records management systems, students as well as education administrators had to make do with manual paper based records and filing systems. Due to their nature these manual filing systems posed tremendous challenges for simple administrative tasks such as searching, aggregation and summarization as well as timely and accurate reporting especially in cases where large volumes of data were involved.

Beginning the early 80s when personal computers (PCs) were introduced, computing power became reasonably affordable for even small educational institutions to invest in computing systems. Following this development, lots of commercial software both for educational institutions and other industries became readily available and affordable. With the introduction of database systems in the late 80s and early 90s, most academic institutions invested in student management systems such that by early 2000 most academic institutions had amassed large volumes of data containing student records of all manner, from financial to academic to e-learning systems resource usage and so on. Other than rudimentary reporting, the next challenge then became how to make sense of, or derive knowledge from these large volumes of educational data.

This is when educational data mining was born. Data mining had already been around as a discipline but had until then been largely applied to problems such as customer relationship management, retail sales and marketing, insurance and similar areas.

Educational data mining is essentially the application of data mining techniques to educational data with the objective of deriving knowledge that can be of benefit to students, education administrators, educational system developers, courseware developers and course content developers. Fundamentally it entails the application of data mining techniques such as classification, clustering, prediction, outlier detection, association rule mining, profiling and so on to educational data. Educational data mining has been used to personalize e-learning systems, to predict student performance, to help students select majors, among many other problems.

1.2 Problem Statement

In modern academic institutions of higher learning, tutors are faced with the problem of large numbers of students, multiple courses to teach in different institutions not to mention the intricacies of prerequisite courses, continuous assessment tests and a host of other examination regulations and requirements. In this situation, the tutor is still required to be able to give individual advice to students on how to achieve the best performance in the shortest time possible taking into consideration previous student performances, examination regulations and various other factors that are likely to affect the student's performance. This also involves helping the students to select an area of specialization in which they are competent in based on their performance.

As trivial as it may seem, this is a very critical problem in institutions of higher learning that needs to be addressed in order to ensure high education standards as well as highly qualified graduates. By addressing this problem, academic institutions will also ensure higher completion rates, greater success of graduates in the job market not to mention improved management of academics in the institutions.

This project proposes the use of data mining techniques to improve the quality of student academic advising. This will be done by developing a data warehouse of the relevant educational data, application of data mining techniques such as statistics and visualization, classification, clustering and outlier detection to help extract and present useful knowledge based on both current and historical data. This knowledge will then be made readily available to both the students, tutors and educational administrators to ensure they are able to provide useful and timely advice to individual students.

1.3 Objectives

The objectives of this project are the following:-

1. To research on the problems in the education sector and the educational data mining techniques which are used to solve these problems
2. To show how a data warehouse suitable for educational data mining can be developed from a relational SMIS

3. To perform data mining on the developed data warehouse to extract knowledge that is useful for student academic advising from the data warehouse

4. To show how a data mining solution can be integrated in an existing SMIS

1.4 Research Questions

Upon completion and successful conclusion, this research aims to answer the following questions:-

1. What are the problems in the education sector that can be solved using educational data mining?

2. Which data mining techniques are applied to the solution of problems in the education sector?

3. Which data mining tools can be employed to educational data?

4. How can a data warehouse of educational data be constructed from a SMIS?

5. Can useful knowledge be derived from educational data for purposes of academic advising?

1.5 Scope

Despite the fact that data mining can be applied to solve problems whose solution is of benefit to lecturers, educational administrators, course content and courseware developers and so on, this project will concentrate on application of educational data mining to provide student academic advisors with factual knowledge that they can use to advise students with the goal of improving student performance, increasing completion rates and improving the overall success of graduates in the job market.

The problem of student academic advising is not only limited to institutions of higher learning but can similarly be applied to other levels of education such as secondary and primary levels. This project will concentrate on application of Educational Data Mining in institutions of higher learning.

Since the inception of data mining, various tools both open source and proprietary have been developed for the development and implementation of data mining solutions. This project

will use Microsoft Business Intelligence Development Studio 2008 and Microsoft SQL Server 2008 to develop a prototype of the proposed solution.

This research project will aim to review the latest developments in the field of education data mining and their application in Kenyan institutions of higher learning. Using a particular private Kenyan university as a case study, a data warehouse and a data mining prototype will be developed to demonstrate how the problem can be solved using educational data mining techniques.

1.6 Assumptions and Limitations

The following are some of the assumptions and limitations that will bind the project:-

- That the student performance data is an accurate representation of the true achievement of individual students and that cases of cheating have been mostly eliminated
- That the student performance data and bio data is largely free of errors

CHAPTER 2

LITERATURE REVIEW

2.0 Introduction

The purpose of this chapter is to provide an overview of the field of Educational Data Mining (EDM), to review the types of problems that EDM attempts to solve, to identify data mining techniques which are applied in EDM, to present an overview of previous research topics in EDM as well as highlight future trends in the field of EDM.

2.0.1 Educational Data Mining

EDM as earlier defined is concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in (Ryan, Yacef, 2009). EDM just like general data mining is concerned with extracting knowledge from large collections of historical data using a subset of the general data mining techniques that have been developed over the years such as statistics and visualization, web mining, clustering, classification, outlier detection, association rule mining, sequential pattern mining, and text mining (Romero, Ventura, 2007). In their survey published in 2007, Romero and Ventura show that substantial research has been done on the application of these data mining techniques to EDM by other researchers in the period between 1995 and 2005.

EDM can be performed to provide knowledge that is of benefit to different categories of players in the education industry i.e. students, instructors, course content developers, learning management system developers and educational administrators to name just a few (Romero, Ventura, 2010). In this paper, Romero and Ventura show that EDM provides immense benefit to all these categories of players in the education sector.

The data that is mined in EDM can come from two types of education systems namely traditional classrooms and distance education (Romero, et al., 2007). This differentiation is important since the data from the two systems requires application of different data mining techniques. For this project, data from the traditional classroom environment will be used mainly student course registrations, student performance data and student demographic data.

2.1 EDM Problems

For a long time, data mining was most popularly applied in the fields of Customer Relationship Management, Sales, Marketing, Insurance and Credit. The problems that were addressed in these applications were mainly increasing sales, customer retention, customer churning, fraud detection, direct marketing, and predictions to name but just a few.

EDM on the other hand attempts to solve problems in educational settings such as analysis and visualization of data, providing feedback for supporting instructors, recommendations for students, predicting students' performance, student modeling, detecting undesirable student behaviors, grouping students, social network analysis, developing concept maps, constructing courseware as well as planning and scheduling (Romero, et al., 2010).

The problems stated by Romero and Ventura in their review of educational data mining 2010, forms the basis of the main problems that are addressed by EDM. These problems are briefly described below:-

i. Analysis and visualization of data – this problem is concerned with the employment of statistics and other analytics techniques to provide graphical views of student activities in learning systems, their performances and so on with the objective of improving student learning and performance (Baepler, Murdoch, 2010). Baepler et al (2010) shows that academic analytics and data mining can be used to provide knowledge for use in course redesign, implementation of new assessment lines and communication between instructors and students.

ii. Providing feedback for supporting instructors – here the challenge is to obtain information from students that can be employed by instructors to improve their teaching and content delivery techniques. Badur and Mardikyan (2011) discuss the use of stepwise regression and decision tree data mining techniques used to analyze the performance of teaching instructors based on student evaluations of the instructors.

iii. Recommendations for students – by employing data mining techniques on large data sets of student interaction with learning systems for example, students can be provided with learning recommendations that help to identify dropouts and students who need special attention (Baradwaj, Pal, 2012). Baradwaj et al (2012) describes

how decision trees can be used to perform classification of students thereby identifying potential drop outs and those who need special attention.

iv. Predicting students' performance – using data mining techniques, student performances can be predicted and the resulting predictions used to counsel the students so as to maximize success and completion rates. Ramaswami and Bhaskaran (2010) use the CHIAD predictive model to identify slow learners and to study the influence of dominant factors on student performance. They extract a set of prediction rules from this model and show that the efficiency of this prediction model is satisfactory.

v. Student modeling – by analyzing data from eLearning systems, student behavior patterns can be identified and studied. An approach that uses clustering to characterize student behavior is employed by Talavera and Gaudioso (2004).

vi. Detecting undesirable student behaviors – these include behaviors such as failure, dropping out and so on which can be identified and corrected in good time. Tair and El-Halees (2012) use classification, clustering and association rule mining to extract knowledge that can be used to improve performance as well as identify potential failures.

vii. Grouping students – similar to student modeling but groups students not based on their mental models but by their customized features and personal characteristics in a learning management system. These groups can then be employed in the development of personalized learning systems that promote effective group learning. Talavera et al (2004) describes how clustering can be used to study collaboration of students in an eLearning system. This knowledge can then be used to group students into categories which can then be used to provide custom interfaces to the eLearning system.

viii. Social network analysis – the interests of students or tutors can be used to provide automatic recommendations to them about others who have similar interests to promote collaboration. Ismail (2012) explains how K-Means clustering can be used to mine tutors interesting areas and thereby assist tutors to find other tutors who have

similar interests. This type of knowledge can serve to promote collaboration among tutors and even simplify the process of academic paper writing.

ix. Developing concept maps – Concept maps are conceptual graphs that show relationships between concepts and express the hierarchical structure of knowledge. Using EDM techniques, concept maps can be auto constructed by employing data mining techniques such as association rule mining and text mining. These concept maps can then be used to overcome the learning barriers of students and their misconceptions (Lee C., Lee G., 2009).

x. Constructing courseware – using knowledge from EDM, better courseware can be constructed and furthermore this can be done automatically. Tang, Lau, Q. Li, Yin, T. Li and Kilis (2000) present an example of how this can be done using web data mining.

xi. Planning and scheduling – this involves activities such as course scheduling, resource allocation, admission and counseling among others. Decision trees and Bayesian models have been proposed to help in probing the effect of changes in recruitment, admissions and courses on education (Ranjan, Khalil, 2008)

This project, will explore the problems of analysis and visualization of data, providing feedback for supporting instructors and students, detecting students likely to fail, and predicting student performance. The solution of these problems will be illustrated using a specific case study of a Kenyan private university.

Data mining is a fast growing field and already new trends are fast coming up that promise to provide even better solutions to data mining problems. Some of these include distributed data mining which if applied in the educational field, means that data mining algorithms can be run on data warehouses across different learning institutions thereby yielding better results.

In the line of EDM, future research areas include development of simpler EDM tools, standardization of methods and data, integration with eLearning systems, and specific EDM techniques (Romero, et al., 2007). This project will seek to demystify EDM by showing how EDM can be performed using a readily available tool i.e. Microsoft Business Intelligence

Development Studio, show how a data mining solution can be developed and integrated to an existing SMIS.

2.2 EDM techniques

Romero and Ventura in their survey of EDM 2007 (Romero, et al., 2007), provide a list of EDM techniques that can be employed in the solution of EDM problems. This list is not exhaustive since over time various other researchers have come up with other techniques that can also be employed in the solution of EDM problems. Some of the most commonly employed EDM techniques according to Romero and Ventura include the following:-

i. Statistics and visualization – this technique employs the use of statistics and data analytics to graphically illustrate educational data from different perspectives as well as providing useful summarizations. This technique is related to the area referred to as learning analytics. (Baepler, et al., 2010)

ii. Web mining – this refers to the application of data mining techniques such as clustering, classification and association rule mining to data obtained from eLearning systems and other web activities involving interaction between students, tutors and resources (Talavera, et al., 2004).

iii. Clustering – this refers to grouping data based on similarities without using predefined classes (Sunita, Aher, Lobo, 2011)

iv. Classification – unlike clustering, this technique groups data based on predefined classes and examples. Brijesh, et al. (2011) shows how this can be accomplished.

v. Outlier detection – refers to application of data mining techniques such as statistics, classification and clustering to detect abnormal occurrences such as student failures. This technique is demonstrated by Umesh and Pal (2011).

vi. Association rule mining – also referred to as affinity grouping, this is a technique that is used to identify things that go together. Ismail (2012) shows an implementation of association rule mining to mine tutor's interesting areas.

vii. Sequential pattern mining – concerned with mining data which occurs as a sorted set of items i.e. the items occur in particular orders and these orders can be studied. Su J.,

Tseng S., et al. (2006) investigate learning portfolio analysis and mining using sequential pattern mining.

viii. Text mining – refers to extraction of patterns from textual data with the objective of classification or summarization. (Ueno M., 2004)

This research project will employ analysis and visualization, classification, clustering, and outlier detection to provide knowledge that can be used by academic advisors to counsel students on how to ensure they perform best.

2.3 Data Mining tools

A variety of data mining tools exist for general data mining which can also be applied to EDM such as DBMiner, Clementine, Intelligent Miner, Weka among others. Some specific EDM tools exist as well such as MultiStar, TADA-ED, 03R, Synergo among others (Romero, et al., 2007). This project will employ Microsoft Business Intelligence Development Studio, Microsoft SQL Server and Microsoft Analysis Services.

2.4 EDM Society and Research

EDM research is currently spearheaded by the International EDM society (www.educationaldatamining.org) which runs the International Journal of Educational Data Mining (JEDM). The society also organizes annual conferences on EDM since 2008, a total of five conferences have been held so far with the sixth one due to be held in July 2013.

Other related organizations that promote research in the same area or related areas include the International Artificial Intelligence in Education Society (IAIED), and the International Society for Learning Sciences (ISLS).

CHAPTER 3
METHODOLOGY

3.0 Introduction

Employment of a data mining solution is usually suitable in situations where ad hoc queries cannot be applied. These situations normally involve for instance complex questions which cannot be easily formulated into queries or cases where the actual knowledge sought is not known beforehand. By running data mining algorithms on appropriate data sets, knowledge can be discovered which would otherwise not be possible using ad hoc queries.

In some instances, the querying and reporting requirements may be so enormous that they would virtually shut down a transactional database. In such a case, it is normally necessary to separate transactional data from reporting data by the construction of a data warehouse optimized for Online Analytical Processing (OLAP).

Several data mining process models have been developed over the years among them Knowledge Discovery in Databases (KDD), Sample Explore Modify Model and Assess (SEMMA), and Cross Industry Process Model for Data Mining (CRISP-DM) (Azevedo, Santos, 2008). These process models provide steps that guide a data mining solution development. Based on a survey conducted in 2007 by Ponce and Karahoca, CRISP-DM was found to be the most popularly used process model (Ponce, Karahoca, 2009).

This project will employ the CRISP-DM process model in the implementation of the proposed data mining solution. The CRISP-DM methodology employs six steps and is illustrated below:-

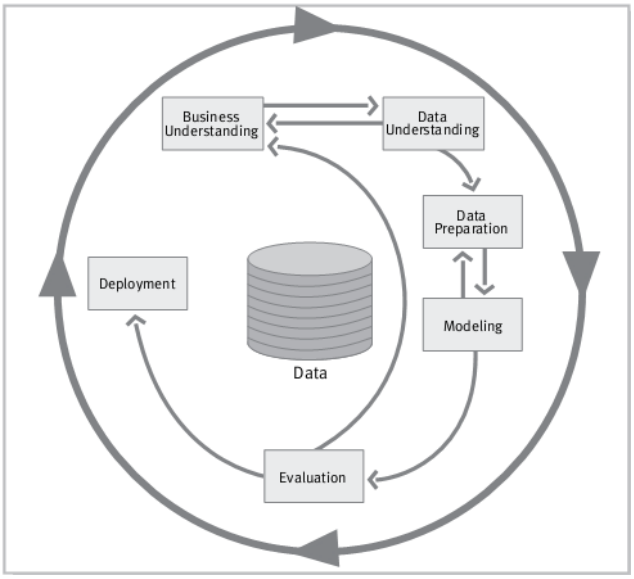

Figure 1: The CRISP-DM Process Model

The steps employed in the CRISP-DM model are described below:-

1. Business understanding – this phase entails understanding the objectives and the requirements of a project, translating the project into a data mining problem and translating of the project objectives into a plan.
2. Data understanding – this entails data collection, identification of data quality problems, and identification of subsets that hypotheses can be formed about.
3. Data preparation – involves extraction and transformation of data into a form that will be loaded into the modeling tool. Also entails selection of the most important attributes.
4. Modeling – this step entails selection of appropriate modeling techniques and their application. Different modeling techniques will normally have specific requirements on the data as a result of which a return to the data preparation stage may be necessary.
5. Evaluation – the model developed is thoroughly evaluated to ensure it captures all the issues of the problem to be solved and achieves all the objectives of the project.
6. Deployment – in this step the results of the data mining process are interpreted and presented using reports.

CRISP-DM as an approach to knowledge discovery in databases or data mining has been defined as a standard by the consortium of companies that formed it Chapman et. al (2000). This standard contains detailed descriptions of all the tasks that are undertaken under each

phase and the corresponding outputs from each phase task. These phases and tasks are briefly highlighted in the table below:-

PHASE	**TASKS**	**Output**
Business Understanding	– Determine business objectives – Assess the situation – Determine data mining goals – Produce project plan	– Business objectives – Business success criteria – Inventory of resources – Requirements, assumptions and constraints – Data mining goals – Project plan
Data Understanding	– Collect initial data – Describe initial data – Explore data – Verify data quality	– Data collection report – Data description report – Data exploration report – Data quality report
Data Preparation	– Select data – Clean data – Construct data – Integrate data – Format data	– Rationale for inclusion/exclusion of data – Data cleaning report – Derived attributes – Generated records – Merged data – Reformatted data
Modeling	– Select modeling techniques – Generate test design – Build model – Assess model	– Modeling techniques – Test design – Parameter settings – Model assessment, revised parameter settings

Evaluation	– Evaluate results – Review process – Determine next steps	– Assessment of results with respect to business success criteria – Approved models – Review of process – List of possible actions
Deployment	– Plan deployment – Plan monitoring and maintenance – Produce final report – Review project	– Deployment plan – Monitoring and maintenance plan – Final report

Table 1: The CRISP-DM Phases, tasks and outputs

Application of CRISP-DM therefore as described in the standard Chapman et. al. (2000) entails identification of applicable phases and tasks to a data mining project, tabulating of these, deletion of unnecessary phases and tasks depending on the project and addition of new tasks under the relevant phases depending on any special requirements of the project.

Since the output of this study is merely a prototype, the last phase of deployment is deemed unnecessary since the prototype is not a fully operational system that will be deployed and maintained. Instead the prototype will only be used to demonstrate how the proposed model can be put to use in order to meet the study objectives.

The actual phases and tasks as per CRISP-DM that will be employed in this study are detailed in the table below:-

PHASE	**TASKS**	**Actual Tasks**	**Actual Outputs**
Business Understanding	– Determine business objectives – Assess the situation – Determine data mining goals – Produce project plan	– The overall objectives of the project – Determine the availability of sufficient resources for the project – These are the specific data mining tasks such as	–

		the prediction, and classification – A project plan which is part of this documentation	
Data Understanding	– Collect initial data – Describe initial data – Explore data – Verify data quality	– Obtain data from the source system(s) – Description of the fact and dimension tables – Create preliminary pivot tables, charts etc to understand the data – Ensure missing/ incorrect data will affect the project to a large degree	–
Data Preparation	– Select data – Clean data – Construct data – Integrate data – Format data	– Determine actual data for the project – Remove unnecessary records and column details – Create the data warehouse – Set appropriate data types and attribute types for the data tables e.g. determine input and predict columns	–
Modelling	– Select modeling techniques – Generate test design – Build model	– Determine the mining algorithms that will be executed – Develop actual data	–

	– Assess model	warehouse – Build a model for each data mining task – Use the relevant techniques to gauge the appropriateness of the models	
Evaluation	– Evaluate results – Review process – Determine next steps	– Use expert reviews to gauge the usefulness of the results – Check the entire process for errors or omissions – Show how the solution can be integrated into an existing system	–

Table 2: Application of CRISP-DM to the Project

None required tasks:-

- Dataset description (Data Preparation)
- Test design (Modeling)
- Model Assessment (Modeling) – will be provided for in the evaluation section

3.1 Data

Successful data mining projects usually employ the use of large collections of data normally subject oriented, historical and time variant. The question therefore becomes how large should the data set be? The goal for successful data mining is always to ensure the data covers all possible phases of change in the subject being investigated.

In order to meet this requirement, this project will aim to use data that covers at least a complete cycle of a student's progress in school i.e. from the first year up until the students complete their course. By adhering to this requirement, the investigation will cover all questions that may arise concerning student academic advising at any stage in the course of their learning.

The sources of data for this project will be student course registrations, student performance grades, student evaluations of tutors, class timetables, as well as student demographic data. This data shall be extracted from all the different sources mainly Microsoft excel and SQL Server databases, transformed into the appropriate formats and loaded into a data warehouse.

3.2 Data Mining Tasks

Once the data has been loaded into a database, the appropriate data mining tasks will be executed. The data mining tasks that will be performed will include:-

1. Classification – classify the students into good performers and underperformers.
2. Clustering – employ statistics to identify natural groupings of students based on performance, demographics, course registrations and the relationships between these.
3. Association – run association rule mining algorithms to extract existing relationships between for example student performances and demographics.
4. Outlier Detection – use statistical measures to determine abnormal student occurrences for instance in terms of performance and the relationship between performance and other student characteristics.
5. Prediction – run data mining algorithms to predict student performances in the final year based on their performances in the first and second years as well as other factors.

3.3 Hardware and Software Requirements

3.3.1 Microsoft SQL Server 2008 R2

Implementation of the data warehouse will require Microsoft SQL Server 2008 database, Microsoft Integration and Analysis Services. The integration services are necessary for performance of data extraction, transformation and loading into the data warehouse. The analysis services engine is necessary for performance of actual data mining and reporting tasks on the data mining process.

3.3.2 Microsoft Business Intelligence Development Studio 2008

In addition, Microsoft Business Intelligence Development Studio 2008 will be required for the implementation and execution of some of the data mining tasks as well as data visualization and reporting. Since some of the data will initially be in excel format,

Microsoft Excel will be very useful in the preprocessing and cleaning of the data before it is loaded into the data warehouse.

3.3.3 Server Computer

For hardware, a server or computer cable of running the 64 bit edition of Windows Server 2008 R2 will be required. This is not strictly necessary since other operating systems such as Windows 7 can run Microsoft SQL Server 2008 save for the fact that the recommended operating system would perform best.

3.4 Requirement Specification and Analysis

3.4.1 Functional Requirements

These are the specific functions that the system will be expected to perform in converting specific inputs to desired outputs through processing.

3.4.2 The Data Warehouse

The data warehouse that will be implemented shall perform the following functions:-

i. Store the data that will be used for the data mining process in a star/snowflake schema from whence it can be efficiently accessed by the data mining algorithms
ii. Present summaries and measures from the data that will be used to answer the data mining questions in the form of a knowledge repository.
iii. Provide security controls for accessing the data and execution of data mining tasks.

3.4.3 Data Warehouse Design

The data warehouse that will be implemented in this study will be based on the Multidimensional Modelling as opposed to the materialized views approach. In this approach, the key concepts include:-

- Star or snowflake design
- Fact tables
- Dimension tables
- Dimension hierarchy
- Measures

The multidimensional approach that is employed is derived from the traditional database design approach which includes four steps:-

1. Requirements analysis – determine the required attributes and the purpose of each as well as typical queries
2. Conceptual design – develop a graphical representation of the schema design showing facts, dimensions and attributes
3. Logical design – convert conceptual schema above into a logical one with respect to the target logical data model
4. Physical design – implement logical schemata in target database system

Process model for conceptual design:-

1. Context definition of measures
2. Dimension hierarchy design
3. Definition of summarizability constraints – since not all aggregations make sense

The data required for the development of the data warehouse in form of fact and dimension tables as well as the function of each field is illustrated in the table below:-

Table Type	**Table Name**	**Table Fields**
Fact	Students	– StudentUID – BirthDate (use to compute age only indicated by youth, middle-age, mature) – ProgramID – CampusID – BillingTypeID – AttendanceTypeID – Gender – BirthCountry – Religion – Marital Status – Employment Status – Fee Payment Time

		Average (First, Second or Last Month) – Degree Class – High School Grade – Average No. of Units per trimester – Average Attendance
Dimension	Campus	– CampusID – Campus Name – Location
Dimension	Program	– ProgramID – Program Name – Department
Dimension	High School	– High School ID – High School Name – Town
Dimension	AttendanceType	– AttendanceTypeID – AttendanceTypeName
Dimension	BillingType	– BillingTypeID – BillingTypeName

Table 3: Fact and Dimension tables

The resulting conceptual schema developed for the data warehouse is a star schema which is illustrated below:-

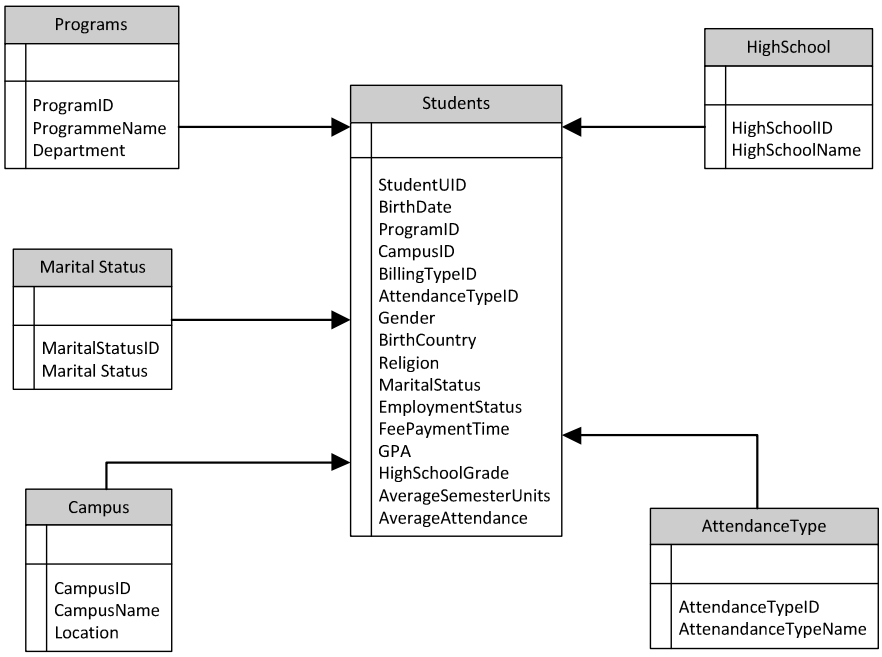

Figure 2: Data Warehouse Conceptual Schema

3.5 Data Analysis and Data Mining

The data mining techniques that will be implemented as part of the project will enable performance of the following:-

i. Prediction of student performance
ii. Identification of potential failures among students
iii. Identification of student natural groupings based on their performance and interpretation of these.
iv. Identify the optimal order of course registration for overall best performance.

3.6 Client Side Interface

The client side of the project will entail integration of the data mining results/recommendations into the tutor/student web portals. This integration will be implemented using asp.net programming language and integrated into the existing learning management system.

3.7 Non-Functional Requirements

i. Capacity – the data warehouse will be able to hold at least 10GB of data from all the sources required to implement the data mining solution proposed by the project.

ii. Response Time – in order for the system integration into student and tutor web portals to be useful, the system response time will be less than 10 seconds.

iii. Throughput – since the system will be queried by approximately 100 users at any particular time, the system should be able to process this number of simultaneous requests within the set time of less than 10 seconds specified above.

iv. Reliability – both tutors and students will be basing their decisions on the output of the system, therefore the data mining results must be accurate at all times.

v. Security – due to the privacy concerns regarding the data that is contained in the database, stringent security controls will be enforced.

3.8 Design

3.8.1 Conceptual Framework

The student academic advising system prototype will made up of a client side interface for both students and staff, extraction, transformation and loading tools, data mining and analysis tools, a knowledge repository as well as a data warehouse that will house all the different types of data. The resulting system will be based on the framework presented by Nguyen, Huang, et al (2008) which is shown below:-

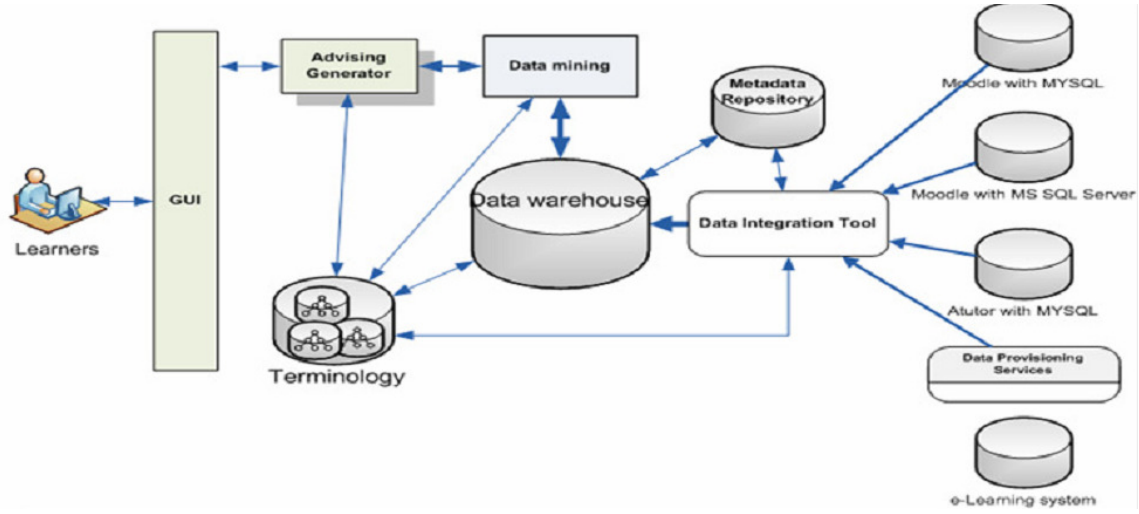

Figure 3: Nguyen, Huang et. al. Conceptual Framework

3.8.2 Modified Conceptual Framework

The modified conceptual framework that is employed in this study is illustrated below:-

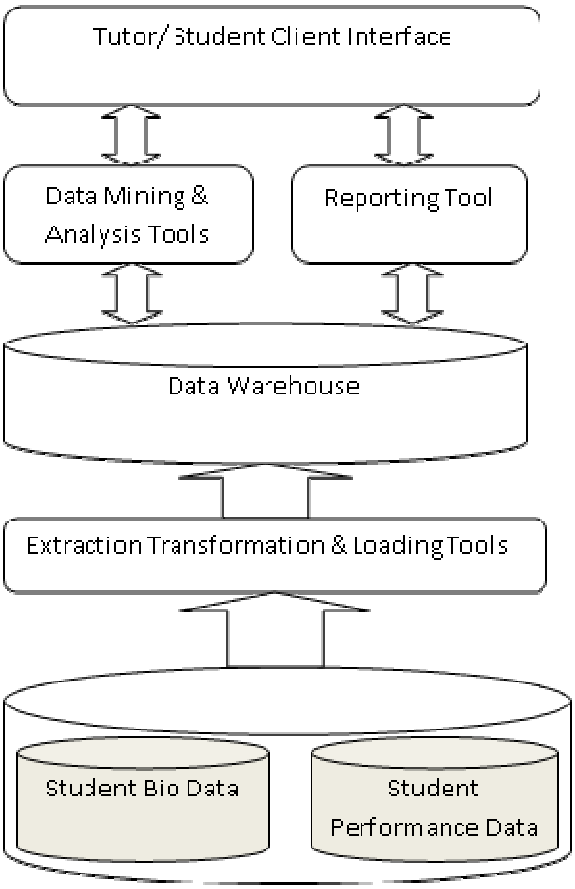

Figure 4: Proposed Conceptual Framework

Figure: Proposed Conceptual Framework for Knowledge Based Student Academic Advising

These components are briefly described below:-

i. Client Side Interfaces – these will enable both the students and the tutors to interact with the system and obtain knowledge that will have been extracted from the data warehouse.

ii. Knowledge Repository – this will contain knowledge extracted from the data warehouse in form of table summaries, and reports.

iii. Data Mining and Analysis Tools – these will be used to interrogate the data warehouse to build knowledge from the data warehouse.

iv. Data warehouse – this will contain the actual data that will have been extracted from different sources, transformed and loaded into star or snowflake schemas that will be mined.

v. Extraction, Transformation and Loading Tools – these tools will be used to extract data from the different sources, convert it into useful forms and load the same into the data warehouse. These tools are also normally used to update the data warehouse periodically.

The extraction, transformation and loading tools are provided by Microsoft SQL Server 2008 and will be used to obtain the data from the original data sources such as excel, and SQL server databases and load these into the data warehouse. The data mining tools used to perform the data mining tasks and provide the necessary results in form of reports and tables marking up the knowledge repository. The client side interfaces will be used to access and employ the data mining results in decision making or actions by both the students and tutors.

3.9 Evaluation and Testing

Two levels of testing will be employed to evaluate the validity of the prototype that will be developed by the study, one to establish the validity of the data mining models developed and another to establish the usefulness of the knowledge extracted from the data warehouse to student academic advising.

3.9.1 Evaluation of the Data Mining Models

The data mining models in the prototype that will be developed in this study will be tested using the following approach:-

1. The entire data available for the project will be split into two, 30% of it will be used to train the models, the rest of the data will be used for testing the models after development and training.
2. The testing techniques that will be employed to verify the validity of the models include the following:-
    a. Using the mining accuracy chart – this testing approach compares the degree of predictive accuracy obtained by comparing a mining model's prediction to the ideal situation (arrived at by using all the available data to train the model) and also comparing it to random guessing.

b. Classification matrix – this entails running the models using all the data and establishing how many times the model gives the correct predictions and how many times it gives incorrect predictions. This measure gives an indication of the model's predictive accuracy.
c. Cross validation – this is a technique used to establish how good the training data for a model is. The training data is split into partitions which are then trained using the data from all the other partitions except that for the specific partition. Each partition specific model is then run using its partition specific data. If each partition specific model performs well then using the entire data set to train a model will result in a good model. For a good model to result, the results of all the partitions must be fairly similar.

3.9.2 Evaluation of the Knowledge Obtained Through the Mining Process

Evaluation of the usefulness of the knowledge obtained to student academic advising shall be accomplished by use of expert reviews. This shall entail demonstration of the system to experienced tutors in the case institution of higher learning and obtaining their views on whether the knowledge obtained is useful for student academic advising. This shall be done by use of checklists that require the respondent to score various aspects of the prototype and provide some general comments. These responses shall be summarized and aggregated to determine the general consensus which will act as an indicator of the validity of the models developed in this study in the real world.

CHAPTER 4.0
IMPLEMENTATION AND FINDINGS

4.0 Introduction

This chapter details the implementation of a prototype of the proposed conceptual framework. The results that were obtained are also presented and discussed as well as the tests for validity of the models implemented in the prototype.

4.1 Data Mining Process

The overall data mining process was governed by the CRISP-DM framework as outlined in the methodology. The objectives/ requirements of the data mining process were taken to be those specified in the study objectives. The data required to achieve these objectives was obtained from the SMIS of the case institution, the data mining models were developed and run using testing data that had been set aside from the overall data. A detailed explanation of the data mining process including model implementation and testing is presented below.

4.2 Business understanding

4.2.1 Background

A tutor in an institution of higher learning in modern day Kenya is faced with the challenge of guiding or advising students on how they can perform best and complete their degree programmes in the shortest period of time. This task however, up until the advent of EDM was largely done based on tutor professional experience or professional hunch. This study proposes to glean useful knowledge from student management information systems that can be used by tutors to provide useful, timely and fact based advice to students to enable them succeed in their academics in the shortest period of time possible.

4.2.2 Objectives

This study proposed to achieve the following data mining objectives as detailed in chapter one of this documentation:-

- To show how a data warehouse suitable for educational data mining can be developed from a relational SMIS

- To perform data mining on the developed data warehouse to extract knowledge that is useful for student academic advising from the data warehouse
- To show how a data mining solution can be integrated in an existing SMIS

4.2.3 Success Criteria

The study performed the analysis and visualization, classification, clustering, prediction and outlier detection data mining tasks. The success criteria for each of the data mining tasks were set as the following:-

- Analytics and Visualization – to probe the data and display it pivot tables, charts and compute statistics that are useful for understanding the data
- Prediction – to predict the student degree class a student is likely to get, using student bio data, and student class attendance
- Classification – to develop a decision tree that can be used to classify the students on the basis of the degree class they are likely to get.
- Clustering – to determine the five key natural groupings of students with regard to performance so that these can be used to understand individual students and thus give them personalized advise based on their natural groupings.

4.2.4 Assumptions and Constraints

The following key assumptions are made in this study:-

- That the student performance data is an accurate representation of the true achievement of individual students and that cases of cheating have been mostly eliminated
- That the student performance data and bio data is largely free of errors

4.2.5 Project Plan

A compressed schedule of the activities of this project is depicted in the in the project plan (Appendix I).

4.3 Data Understanding

4.3.1 Data Collection

The data required for this study was obtained from the ANU SMIS and corroborated with the actual graduation lists to ensure data quality. This data was mostly in a relational database system save for the actual graduation lists which were mostly in Microsoft Excel and Word documents. This data was selected and merged using suitable SQL queries, multiple views were also used to understand and merge the data to form the final fact and dimension tables required for the data warehouse. The final tables were imported into the data warehouse schema using SSIS a data extraction, transformation and loading tool provided by Microsoft. This tool was also used to import and convert the data that was in Excel and Word form into the data warehouse.

4.3.2 Data Description

A brief description of the data that was obtained is given below:-

- Student bio data – this data comprises of student personal details, their degree programmes, campuses, birth countries, and attendance type whether regular, distance or evening
- Student performance data – this comprises of student scores in each of the courses taken to complete a degree programme, these are aggregate and an average GPA computed to determine the student degree class
- Student attendance data – this is an approximate record of student class attendance as recorded by the course tutors, the total class attendance for all taught courses was computed and used to inform the models

4.3.3 Exploratory Analysis

In order to understand the data, some preliminary exploration was done. This exploration established the following facts about the data:-

Item	**Facts Established**
Total students graduated (1998-2012)	2084
Majority student birth country	Kenya

Number of Degree Programmes	48
Student average attendance	126

Table 4: Overall Data Summary

Gender	**Population**
Male	605
Female	886

Table 5: Student Population by Gender

Program	**No. of Students**
BACHELOR OF ARTS IN CHRISTIAN MINISTRIES	5
BACHELOR OF BUSINESS AND INFORMATION TECHNOLOGY	170
BACHELOR OF COMMERCE	710
BACHELOR OF DRYLAND NATURAL RESOURCE MANAGEMENT	13
BACHELOR OF EDUCATION - EARLY CHILDHOOD	163
BACHELOR OF EDUCATION - PRIMARY	29
BACHELOR OF EDUCATION - SECONDARY	26
BACHELOR OF EDUCATION - SPECIAL NEEDS ECD	2
BACHELOR OF EDUCATION - SPECIAL NEEDS PRI	78
BACHELOR OF MASS COMMUNICATION	108
BACHELOR OF SCIENCE - COMPUTER SCIENCE	137
BACHELOR OF SCIENCE IN INTERNATIONAL BUSINESS MANA	5
BACHELOR OF THEOLOGY	34
MASTERS OF ARTS IN RELIGION	2
MASTERS OF BUSINESS ADMINISTRATION	9

Table 6: Student Population by Degree Programme

Degree Class	**No. of Students**
First Class	62

Second Lower	620
Second Upper	600
Pass	209

Table 7: Student Degree by Class

4.3.4 Data Quality

In order to establish the validity of the data employed in the study, the data was compared to the actual annual graduation lists as per the graduation booklets. The records which could not be corroborated against the official graduation lists were dropped. The exact details of the number records dropped are contained in the section below: Rationale for Inclusion/ Exclusion.

4.4 Data Preparation

4.4.1 Rationale for Inclusion/ Exclusion

The conceptual schema in the methodology section shows the tables that made up the data warehouse schema. Below is an explanation of the role of each field as well as an explanation of why it was included in the schema:-

Table	Field	Role	Reason for inclusion
Students	Student uid	Primary Key	Uniquely identifies each student
	Age	Input	Measure effect of age on performance
	ProgramID	Input	Compare performances in different programmes
	CampusID	Input	Measure impact of campus on performance
	AttendanceTypeID	Input	Measure variation of performance with attendance type
	Gender	Input	Measure variation of performance with Gender
	BirthCountry	Input	Measure variation of performance with BirthCountry

	Religion	Input	Measure variation of performance with Religion
	MaritalStatus	Input	Measure variation of performance with Marital Status
	EmploymentStatus	Input	Measure variation of performance with EmploymentStatus
	FeePaymentTime	Input	Measure impact of financial challenges on performance
	Degree Class	Predict	This field would be predicted using all the other fields
	High School Grade	Input	Measure variation of performance with high school grade
	Average no. of units per trimester	Input	Measure variation of performance with average no. of units per trimester
	Average Attendance	Input	Measure variation of performance with attendance

Table 8: Rational for Inclusion/ Exclusion of Attributes

The rest of the tables that comprised the data warehouse were dimension tables which were linked to the students' fact table through foreign keys. The dimension tables that were included and the rationale for their inclusion is given in the table below:-

Dimension Table	**Foreign Key**	**Rationale for Inclusion**
Student_Program	ProgramID	Analysis of Students by Degree Program
Student_Campus	CampusID	Analysis of Students by Campus
Student_AttendanceType	AttendanceTypeID	Analysis of Students by Attendance Type
Student_High_School	HighSchoolID	Analysis of Students by High School

Table 9: Dimension Tables

Out of the total population of 12,877 which includes both current and alumni students, only records of 2048 graduands could be corroborated with graduation lists. Of the 2048 graduands, 1802 had their records in electronic form in the SMIS. In order to guarantee accuracy of the models, this number was further reduced to 1491 who had their attendance indicated in the SMIS.

4.4.2 Data Cleaning

In order to ensure the data was in the best possible state to support the data mining objectives, the following was done:-

- Only students who actually completed their programmes and successfully graduated were included
- Several fields in the original data which were empty or were deemed not relevant for the study such as student names, names of their guardians, cell phone numbers and all other contact information were removed
- Duplicate student records were removed
- Primary keys were set for every resulting table and referential integrity enforced. This applied to the fact table, the four dimension tables as well as their related tables.

4.4.3 Derived Attributes and Records

Since the raw data consisted of student performances per course, all these had to be aggregated and the resulting average GPA be used to determine the degree classification for each student. The ages of the students also had to be calculated by obtaining the difference between the birthdate and the reporting date. Finally to get an indication of student class attendance, a count of the student attendance was obtained.

4.4.4 Reformatted Data

The raw average GPA per student would not serve the prediction process very well instead this was converted to be the degree class of the resulting GPA average score. The resulting student performance data therefore only indicated a student degree class as either First Class, Second Class or Pass not the actual average GPA score. This was useful especially in obtaining permission to use the student data since this preserved confidentiality and only used publicly available details in the form of graduation lists.

4.5 Modeling

4.5.1 Data Warehouse

A data warehouse was developed using Microsoft SSAS and SQL Server database. The conceptual schema that was used to implement this is shown in the methodology section, the rationale for inclusion/exclusion of different fields is explained above in the section data preparation.

4.5.2 Modeling Techniques

The data mining tasks that were to be performed included the following:-

- Classification
- Clustering
- Prediction
- Analytics and Visualization

The table below shows the data mining tasks and the corresponding algorithms that were used to perform them:-

Data Mining Task	**Selected Algorithm**
Classification	Microsoft Decision Trees
Clustering and Outlier Detection	Microsoft Clustering
Prediction	Microsoft Naïve Bayes
Analysis and Visualization	This was to be accomplished using Microsoft Reporting Services, a companion tool to Analysis Services the Microsoft data mining tool that was employed in the study

Table 10: Data Mining Tasks and Selected Algorithms

4.5.3 Models

Using Microsoft Visual Studio 2008, a mining structure was defined and the three models for each of the data mining tasks developed. For each of the data mining models, the field settings were set as per the table in the data preparation phase, the models were also all configured to set aside 30% of the data for testing purposes.

The table below gives a brief description of the models and their purposes:-

Model	Algorithm	Purpose
Classification Model	Microsoft Naïve Bayes	This model was used to perform classification of the data in order to provide a deeper understanding of how the various attributes impact performance
Clustering and Outlier Detection Model	Microsoft Clustering	This model was used to group the data based on natural groupings with respect to performance. This model was also used to understand student outliers
Prediction Model	Microsoft Decision Trees	This model was trained and used to predict student performance based on the attributes provided in the fact table

Table 11: Implemented Data Mining Models

4.6 Evaluation and Testing

Each of the data mining models built as part of the prototype was evaluated separately using the four techniques specified in the methodology. The corresponding lift chart, classification matrix and cross validation table was obtained for each of the models and used to assess the models. The following techniques were employed to ensure validity of the models developed and hence the knowledge extracted during the entire mining process:-

- Partitioning data into testing and training sets – 30% of the data was set aside from the onset to be used later for testing the performance of the models in terms of accuracy and validity

- Measuring lift and gain -. a lift chart is a method of visualizing the improvement that you get from using a data mining model, when you compare it to random guessing and an ideal model obtained by training the model using all the data available
- Performing cross-validation of data sets – this model validation technique entails partitioning the data, developing models using all the data
- Generating classification matrices - these matrices sort good and bad guesses into a table so that you can quickly and easily gauge how accurately the model predicts the target value. It shows the total number of correct and incorrect predictions

Below is a detailed illustration and explanation of how each model was assessed.

4.6.1 Lift Charts

The lift charts for all the models are shown below for prediction of Second Upper:-

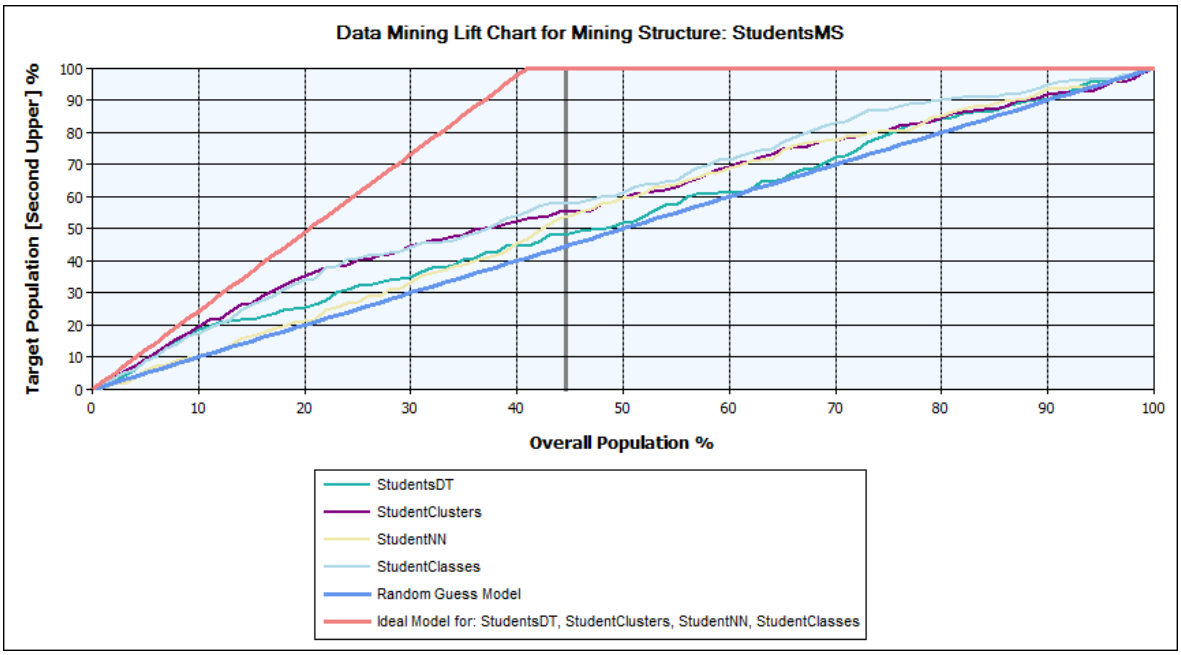

Figure 5: Lift Charts for Predicting Second Upper

For a model to be useful, it should provide a significant degree of lift from the situation that would be achieved if random guessing were used. The upper limit to a model's performance is an ideally performing model which is obtained by training the model using all the data.

As can be seen from the lift charts for each of the models, all the models are significantly above the random line which indicates they perform very well in predicting if a student will obtain a Pass degree class. Due to the small number of First Class degrees, the models

developed do not perform very well in predicting if a student will obtain a First Class degree as shown in the lift charts below:-

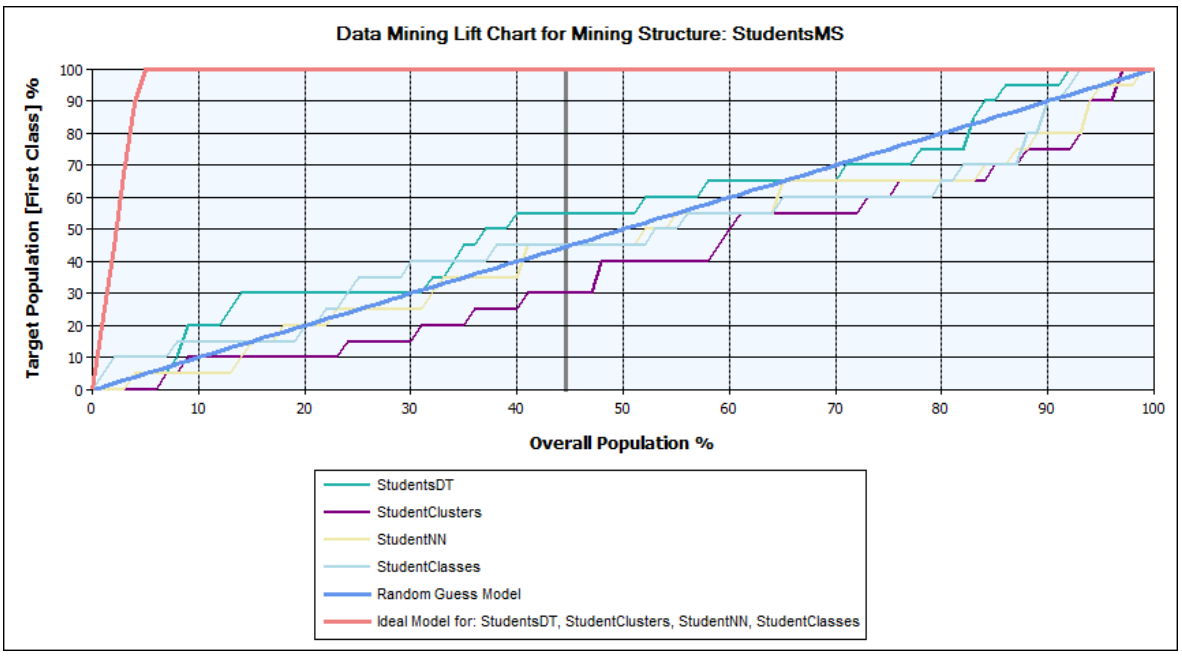

Figure 6: Lift Charts for Predicting First Class

The poor performance of the models in predicting if a student will obtain a First Class degree need not be a problem since this can be accurately deduced by considering the probability that they will get any of the other degree classes.

A comparison of the performance of each of the models can also be obtained by using the a mining legend which shows the percentage population that was predicted correctly and the probability threshold that was required to include a prediction in the correct predictions. For prediction of whether a student will obtain a Pass degree, the comparison mining legend is shown below:-

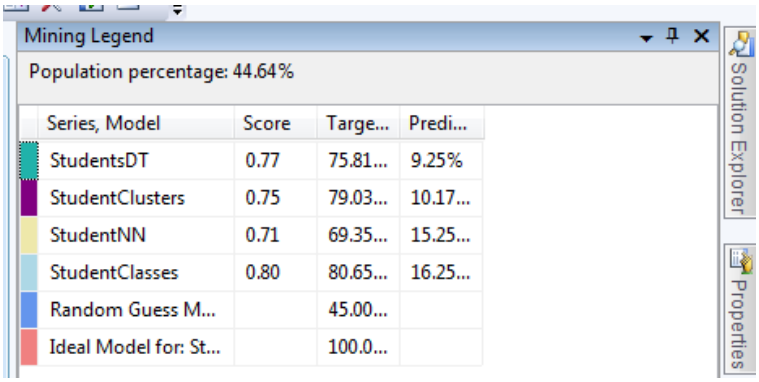

Figure 7: Mining Legend for Ranking Models in Predicting Pass

This mining legend shows that using about 44% of the test cases, the percentage population for which correct predictions were made is shown in the Target column. According to the scores of each model, it can be seen that the clustering model performed best.

4.6.2 Classification Matrices

The classification matrix shows the total number of correctly predicted values as opposed to incorrect predictions. The classification matrix for all the models that were developed are shown below:-

Counts for StudentsDT on Degree Class				
Predicted	First Class (Actual)	Second Lower (Actual)	Second Upper (Actual)	Pass (Actual)
First Class	0	0	0	0
Second Lower	8	113	72	71
Second Upper	12	69	111	43
Pass	0	0	0	0
Total Correct Predictions			224	
Total Incorrect Predictions			141	
Counts for StudentClusters on Degree Class				
Predicted	First Class (Actual)	Second Lower (Actual)	Second Upper (Actual)	Pass (Actual)
First Class	0	0	0	0
Second Lower	17	155	113	59
Second Upper	3	27	70	3
Pass	0	0	0	0
Total Correct Predictions			225	
Total Incorrect Predictions			140	
Counts for StudentNN on Degree Class				

Predicted	First Class (Actual)	Second Lower (Actual)	Second Upper (Actual)	Pass (Actual)
First Class	2	11	0	0
Second Lower	12	131	92	42
Second Upper	8	45	75	16
Pass	0	4	5	4
Total Correct Predictions			212	
Total Incorrect Predictions			137	
Counts for StudentClasses on Degree Class				
Predicted	First Class (Actual)	Second Lower (Actual)	Second Upper (Actual)	Pass (Actual)
First Class	0	2	0	0
Second Lower	13	130	102	55
Second Upper	7	50	81	5
Pass	0	0	0	2
Total Correct Predictions			211	
Total Incorrect Predictions			152	

Table 12: Classification Matrices for all the Models

From these tables it can be seen that the neural network model performs best since it has the smallest number of incorrect predictions. The rest of the models predict correctly for at least more than 50% of the test population.

4.6.3 Cross Validation

Cross validation is a technique of verifying the fitness of the model's training data. Cross validation divides the data into several partitions and then builds models specific to each of the resulting partitions using all the data except that for the specific partition. The partition specific models are then tested using partition specific data. If the accuracy of the partition models is generally good then the overall model will also most likely be good. If the partition

model performance results of the partition models are very similar then the available data is good for the training tasks otherwise the available training data is not enough.

Ten partitions were set and trained as necessary to gauge the suitability of the available data for training the model. The performance of these partition models is summarized in the table below:-

StudentsDT				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	103	Classification	Pass	46
2	103	Classification	Pass	44
3	104	Classification	Pass	51
4	104	Classification	Pass	53
5	105	Classification	Pass	47
6	105	Classification	Pass	49
7	105	Classification	Pass	50
8	105	Classification	Pass	47
9	105	Classification	Pass	37
10	105	Classification	Pass	45
			Average	46.89750958
			Standard Deviation	4.23360411
1	103	Classification	Fail	57
2	103	Classification	Fail	59
3	104	Classification	Fail	53
4	104	Classification	Fail	51
5	105	Classification	Fail	58
6	105	Classification	Fail	56
7	105	Classification	Fail	55
8	105	Classification	Fail	58
9	105	Classification	Fail	68
10	105	Classification	Fail	60
			Average	57.50862069
			Standard Deviation	4.369754372
1	103	Likelihood	Log Score	-1.14776475
2	103	Likelihood	Log Score	-1.11427267
3	104	Likelihood	Log Score	-1.1056338
4	104	Likelihood	Log Score	-1.0905668
5	105	Likelihood	Log Score	-1.10694439
6	105	Likelihood	Log Score	-1.10345166
7	105	Likelihood	Log Score	-1.0946723
8	105	Likelihood	Log Score	-1.11610791
9	105	Likelihood	Log Score	-1.16283905
10	105	Likelihood	Log Score	-1.12167036

			Average	-1.11637137
			Standard Deviation	0.021646315
1	103	Likelihood	Lift	-0.00967644
2	103	Likelihood	Lift	0.035193676
3	104	Likelihood	Lift	0.041114876
4	104	Likelihood	Lift	0.034800564
5	105	Likelihood	Lift	0.026516929
6	105	Likelihood	Lift	0.03000966
7	105	Likelihood	Lift	0.038789014
8	105	Likelihood	Lift	0.017353408
9	105	Likelihood	Lift	-0.02937773
10	105	Likelihood	Lift	0.011790955
			Average	0.019642831
			Standard Deviation	0.021870613
1	103	Likelihood	Root Mean Square Error	0.551811159
2	103	Likelihood	Root Mean Square Error	0.539035512
3	104	Likelihood	Root Mean Square Error	0.545446222
4	104	Likelihood	Root Mean Square Error	0.548403373
5	105	Likelihood	Root Mean Square Error	0.547297972
6	105	Likelihood	Root Mean Square Error	0.53791447
7	105	Likelihood	Root Mean Square Error	0.551980801
8	105	Likelihood	Root Mean Square Error	0.533871709
9	105	Likelihood	Root Mean Square Error	0.541304017
10	105	Likelihood	Root Mean Square Error	0.561731861
			Average	0.545879456
			Standard Deviation	0.007807303
StudentClusters				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	103	Classification	Pass	50
2	103	Classification	Pass	50
3	104	Classification	Pass	49
4	104	Classification	Pass	50
5	105	Classification	Pass	56
6	105	Classification	Pass	54
7	105	Classification	Pass	51
8	105	Classification	Pass	52
9	105	Classification	Pass	50
10	105	Classification	Pass	49
			Average	51.10727969
			Standard Deviation	2.169534944
1	103	Classification	Fail	53
2	103	Classification	Fail	53
3	104	Classification	Fail	55

4	104	Classification	Fail	54
5	105	Classification	Fail	49
6	105	Classification	Fail	51
7	105	Classification	Fail	54
8	105	Classification	Fail	53
9	105	Classification	Fail	55
10	105	Classification	Fail	56
			Average	53.29885057
			Standard Deviation	1.956607486
1	103	Likelihood	Log Score	-1.02520729
2	103	Likelihood	Log Score	-1.08038522
3	104	Likelihood	Log Score	-1.10222179
4	104	Likelihood	Log Score	-1.07126012
5	105	Likelihood	Log Score	-1.04023047
6	105	Likelihood	Log Score	-1.06126121
7	105	Likelihood	Log Score	-1.08446267
8	105	Likelihood	Log Score	-1.0570659
9	105	Likelihood	Log Score	-1.0521673
10	105	Likelihood	Log Score	-1.05328042
			Average	-1.06274644
			Standard Deviation	0.021315075
1	103	Likelihood	Lift	0.112881013
2	103	Likelihood	Lift	0.069081126
3	104	Likelihood	Lift	0.04452689
4	104	Likelihood	Lift	0.054107238
5	105	Likelihood	Lift	0.093230846
6	105	Likelihood	Lift	0.072200109
7	105	Likelihood	Lift	0.048998645
8	105	Likelihood	Lift	0.076395415
9	105	Likelihood	Lift	0.081294017
10	105	Likelihood	Lift	0.080180894
			Average	0.07326776
			Standard Deviation	0.019675069
1	103	Likelihood	Root Mean Square Error	0.515003666
2	103	Likelihood	Root Mean Square Error	0.534966478
3	104	Likelihood	Root Mean Square Error	0.521890905
4	104	Likelihood	Root Mean Square Error	0.528053128
5	105	Likelihood	Root Mean Square Error	0.507846245
6	105	Likelihood	Root Mean Square Error	0.503109655
7	105	Likelihood	Root Mean Square Error	0.515007656
8	105	Likelihood	Root Mean Square Error	0.527792149
9	105	Likelihood	Root Mean Square Error	0.529134521
10	105	Likelihood	Root Mean Square Error	0.518320291
			Average	0.520084491

			Standard Deviation	0.009602557
StudentNN				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	103	Classification	Pass	45
2	103	Classification	Pass	42
3	104	Classification	Pass	53
4	104	Classification	Pass	46
5	105	Classification	Pass	48
6	105	Classification	Pass	40
7	105	Classification	Pass	44
8	105	Classification	Pass	38
9	105	Classification	Pass	42
10	105	Classification	Pass	45
			Average	44.29310345
			Standard Deviation	4.027005384
1	103	Classification	Fail	58
2	103	Classification	Fail	61
3	104	Classification	Fail	51
4	104	Classification	Fail	58
5	105	Classification	Fail	57
6	105	Classification	Fail	65
7	105	Classification	Fail	61
8	105	Classification	Fail	67
9	105	Classification	Fail	63
10	105	Classification	Fail	60
			Average	60.11302682
			Standard Deviation	4.278013696
1	103	Likelihood	Log Score	-1.14766108
2	103	Likelihood	Log Score	-1.11276742
3	104	Likelihood	Log Score	-1.16219268
4	104	Likelihood	Log Score	-1.1689526
5	105	Likelihood	Log Score	-1.17648781
6	105	Likelihood	Log Score	-1.22146143
7	105	Likelihood	Log Score	-1.20155778
8	105	Likelihood	Log Score	-1.25373776
9	105	Likelihood	Log Score	-1.22040882
10	105	Likelihood	Log Score	-1.17347381
			Average	-1.18411075
			Standard Deviation	0.038887286
1	103	Likelihood	Lift	-0.00957277
2	103	Likelihood	Lift	0.036698933
3	104	Likelihood	Lift	-0.015444
4	104	Likelihood	Lift	-0.04358524

5	105	Likelihood	Lift	-0.04302649
6	105	Likelihood	Lift	-0.08800011
7	105	Likelihood	Lift	-0.06809647
8	105	Likelihood	Lift	-0.12027644
9	105	Likelihood	Lift	-0.0869475
10	105	Likelihood	Lift	-0.04001249
			Average	-0.04809655
			Standard Deviation	0.043004281
1	103	Likelihood	Root Mean Square Error	0.496683571
2	103	Likelihood	Root Mean Square Error	0.485020525
3	104	Likelihood	Root Mean Square Error	0.515052471
4	104	Likelihood	Root Mean Square Error	0.488641111
5	105	Likelihood	Root Mean Square Error	0.500291445
6	105	Likelihood	Root Mean Square Error	0.503633377
7	105	Likelihood	Root Mean Square Error	0.50628038
8	105	Likelihood	Root Mean Square Error	0.508381146
9	105	Likelihood	Root Mean Square Error	0.481910413
10	105	Likelihood	Root Mean Square Error	0.474601726
			Average	0.496058425
			Standard Deviation	0.012380141
StudentClasses				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	103	Classification	Pass	52
2	103	Classification	Pass	47
3	104	Classification	Pass	51
4	104	Classification	Pass	52
5	105	Classification	Pass	59
6	105	Classification	Pass	52
7	105	Classification	Pass	54
8	105	Classification	Pass	51
9	105	Classification	Pass	50
10	105	Classification	Pass	56
			Average	52.41283525
			Standard Deviation	3.136588552
1	103	Classification	Fail	51
2	103	Classification	Fail	56
3	104	Classification	Fail	53
4	104	Classification	Fail	52
5	105	Classification	Fail	46
6	105	Classification	Fail	53
7	105	Classification	Fail	51
8	105	Classification	Fail	54
9	105	Classification	Fail	55

10	105	Classification	Fail	49
			Average	51.99329502
			Standard Deviation	2.794862738
1	103	Likelihood	Log Score	-1.02906074
2	103	Likelihood	Log Score	-1.19331133
3	104	Likelihood	Log Score	-1.17134806
4	104	Likelihood	Log Score	-1.14566905
5	105	Likelihood	Log Score	-1.11520085
6	105	Likelihood	Log Score	-1.1739372
7	105	Likelihood	Log Score	-1.10214584
8	105	Likelihood	Log Score	-1.22564979
9	105	Likelihood	Log Score	-1.1181865
10	105	Likelihood	Log Score	-1.04942331
			Average	-1.13242449
			Standard Deviation	0.058914755
1	103	Likelihood	Lift	0.10902757
2	103	Likelihood	Lift	-0.04384498
3	104	Likelihood	Lift	-0.02459938
4	104	Likelihood	Lift	-0.02030169
5	105	Likelihood	Lift	0.018260465
6	105	Likelihood	Lift	-0.04047588
7	105	Likelihood	Lift	0.031315475
8	105	Likelihood	Lift	-0.09218847
9	105	Likelihood	Lift	0.015274815
10	105	Likelihood	Lift	0.084038007
			Average	0.003589712
			Standard Deviation	0.057760974
1	103	Likelihood	Root Mean Square Error	0.486632832
2	103	Likelihood	Root Mean Square Error	0.497243042
3	104	Likelihood	Root Mean Square Error	0.480080645
4	104	Likelihood	Root Mean Square Error	0.498079411
5	105	Likelihood	Root Mean Square Error	0.483212148
6	105	Likelihood	Root Mean Square Error	0.473413267
7	105	Likelihood	Root Mean Square Error	0.493431171
8	105	Likelihood	Root Mean Square Error	0.480219344
9	105	Likelihood	Root Mean Square Error	0.493043072
10	105	Likelihood	Root Mean Square Error	0.494376848
			Average	0.487955867
			Standard Deviation	0.008024183

Table 13: Cross Validation Results for all the Models

From the above tests performed on the data using all the models, it can be seen that the data is indeed suitable for training of the models since the different metrics computed for each partition for each model were very similar.

4.7 User interfaces and Reports

In order for the prediction models developed in this study to be used, one would need to write prediction queries model based queries and build model based reports. To facilitate this Microsoft provides a special querying language called Data Mining Extensions (MDX) for querying the models and a reporting services server application called Microsoft SQL Reporting Services. To illustrate how possible web interfaces can be developed, the study implemented a simple querying and reporting web based interface that can be used by tutors to obtain knowledge for student academic advising. A snapshot of this interface is shown below:-

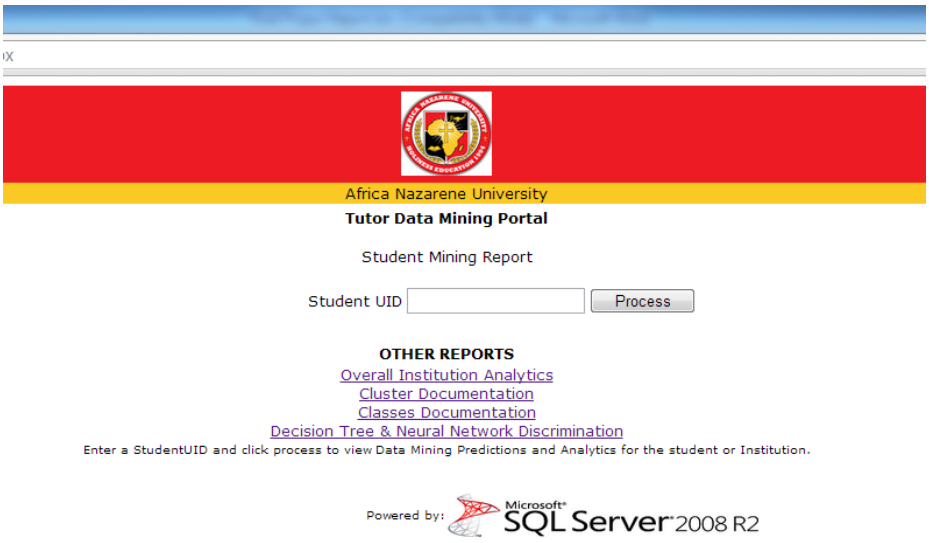

Figure 8: Sample User Interface for Integrating the Proposed Solution in a Web Based Application

To use this interface all that a tutor has to do is specify a student's registration number and then click process. Clicking process queries the data mining models and produces analysis and visualization information that can be used by the tutor to advise the student. Below is a sample report that can be obtained from the system.

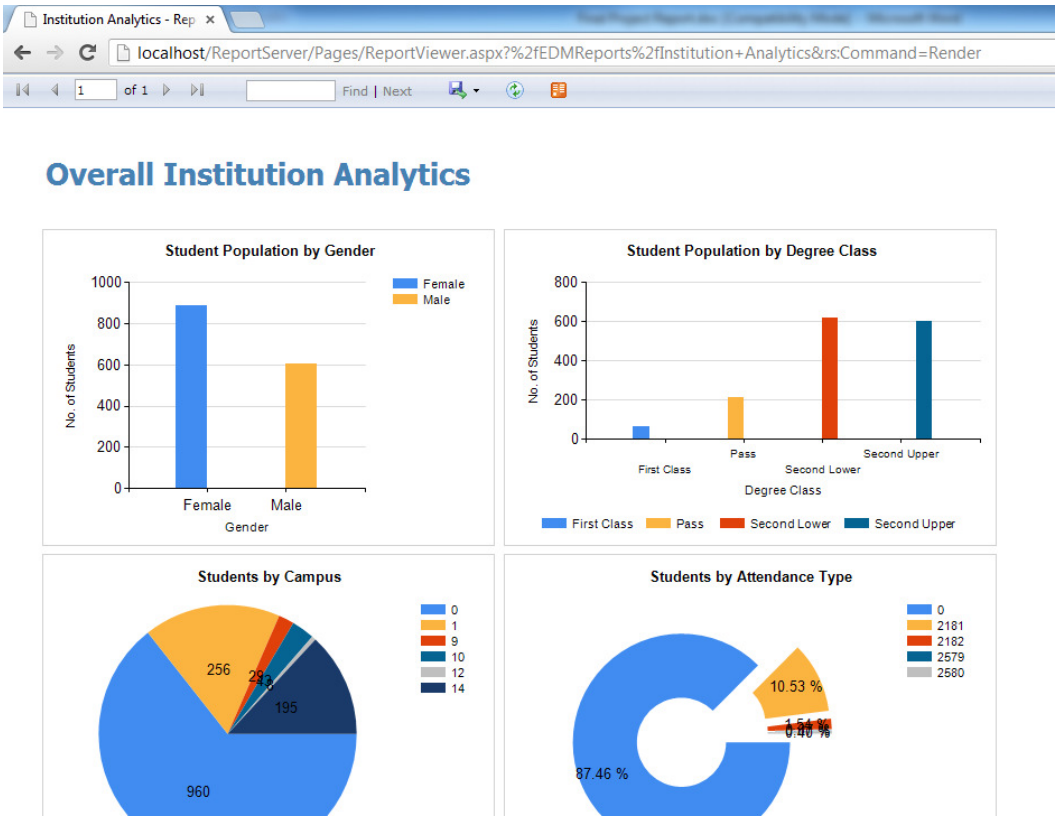

Figure 9: Sample Analytics and Visualization that can be Generated from the Data Warehouse

The interface shows a sample analysis and visualization report as well as the results of querying for the prediction of a student whose final degree class is unknown. The class and cluster of a particular student are also shown.

4.8 Results

4.8.1 Comparison of Data Mining Results to Data Mining Objectives

Analytics and visualization, classification, clustering and prediction were applied to data with the objective of obtaining knowledge that can be used by tutors to advise students. Below is a summary of the knowledge extracted from the data and how it can be applied to the task of student academic advising:-

Knowledge Acquired	**Application**
Analytics and Visualization	The presentation of student analytics, helps the tutor to understand the background against which a student is expected to

	perform. Such statistics as performance by gender, by program and variances in performance help the tutor to put their advice to the student into perspective.
Student Classification	The student classes identified can be used to direct a student on how to study so as to overcome the prejudices of the class they belong to.
Student Clusters	Student clusters point out the characteristics of other students a particular student associates with mostly. These provide an understanding of a student's psyche and can be used guide or counsel the student.
Student Prediction	With a probability indicator of student's likelihood to score a given degree class, students likely to perform very well can be encouraged to continue in the same direction while those likely to underperform can given advise on how to improve.

Table 14: Knowledge Acquired using Data Mining on the Developed Data Warehouse

4.8.2 Expert Review

In order to establish the validity of the results in the real world, the resulting prototype was demonstrated to a head of an academic department, 3 lecturers and a Deputy head of department in the case institution.

The respondents were required to score various aspects of the developed data mining models and their applicability to the task of student academic advising out of 5. Below is a summary of their responses:-

Aspect 1	**R1**	**R2**	**R3**	**R4**	**R5**	**Average**
Are the analysis and visualization results useful for understanding student performance?	4	3	5	4	3	3.6
Are the identified student clusters and classes	4	5	5	4	5	4.8

representative of students in the institution?						
Are the prediction results valid?	3	2	3	3	4	3.2
Are the prediction results useful?	4	3	2	3	2	2.6
Give a score on the overall usefulness of the system and prediction models.	4	3	4	3	4	3.6
Overall Score						3.56

Table 15: Expert Review of the Prototype

Based on the above data, generally the respondents agreed that the models were indeed valid based on their experience and knowledge of the domain. They also were impressed by the knowledge produced by the data mining models and they believe it is indeed useful for student academic advising. Therefore it can be concluded based on their views that the models developed in this study are valid and reliable in addressing the identified problem which is to use knowledge obtained using data mining on educational data to advise students.

4.9 Review of the Process

The results of the data mining process can be improved upon by using slightly more student records to train and implement the models. This would result in significantly more representative models and can be achieved over time by updating the data warehouse with fresh graduation data annually.

The accuracy of the models can also be improved by ensuring that the necessary attributes required for the data mining process are populated for all the students. New attributes can also be introduced based on studies of factors that affect student performance to further improve the accuracy of the prediction models.

Finally feedback obtained from tutors on the application and use of the models to the actual task of student academic advising can be collected and used to improve the models. The results can also be interpreted and presented in a simpler way to understand such that students themselves can also access and use the results to understand what they need to do.

4.10 Resulting Model for EDM

Based on this study, it can be seen that reasonably accurate predictions of student performances can be made using only student bio data and their class attendance. This is data that most institutions already have and already begin to use for the purpose of student academic advising. As a result of this it is not mandatory for special systems and pedagogical data to be obtained for data mining to be employed in educational institutions. To this end, this study proposes the following model which uses only student bio data, attendance and graduation information to perform data mining.

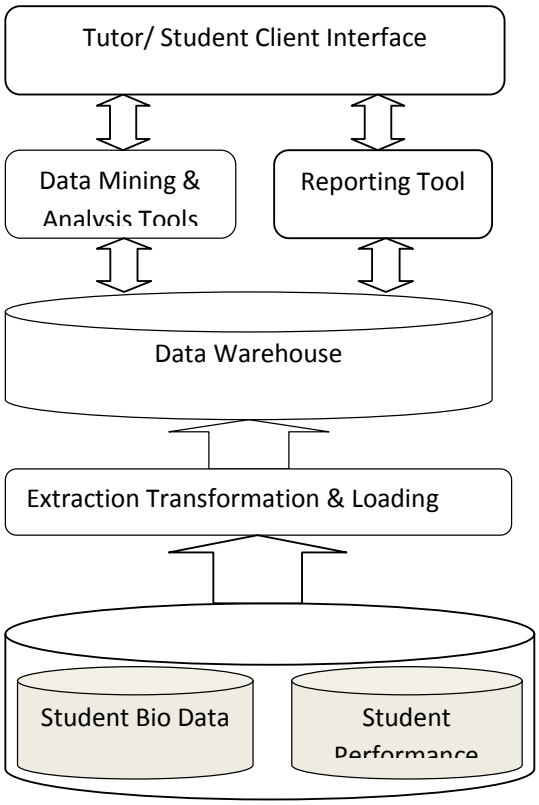

Figure 10: Final EDM Model

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.0 Introduction

Data mining has been around for some time yet its application to the education sector especially in African countries is yet to reach widespread use. A summary of the specific contributions of this study towards widespread adoption of EDM is presented here.

5.1 Contribution of the Project

Early applications of data mining were mainly in the fields of sales, marketing and customer relationship management however, as evidenced by the survey conducted by Romero and Ventura (Romero and Ventura 2007) data mining has also been largely applied in the education sector giving rise to the field of EDM. Using readily available tools i.e. Microsoft Visual Studio and SQL Server this research has shown that the development of a data mining solution and its integration into an existing system is no longer a complex task that is the reserve of experts but rather a simple process that can be executed by even novice data miners.

In addition to rudimentary reports and queries, it has been shown that prediction models can be developed that can predict the final degree honors a student is likely to obtain. Using these predictions together with the information provided by the clusters and classes mined from the data, students can be furnished with advice to enable them achieve the highest honors depending on their abilities. The data that is available to a data mining process is very crucial in that it determines the accuracy of the final results. For instance in this research, the models that were developed were not very good at predicting if a student would score a first class degree due to the small number of first class instances available for training the models. Missing attributes for some of the records were also seen to significantly reduce the accuracy of the models.

Finally it is true that data on several factors that have been identified which affect student performance are not mandatory for one to get started developing EDM solutions. Readily available student bio data, attendance data and even fee payment information is sufficient for the development of reasonably accurate data mining solutions as evidenced in this research.

5.2 Challenges and Limitations

In the course of this study, a number of challenges and limitations were encountered. These are mentioned briefly below:-

- Some attributes were not available for quite a number of students. For example about 300 students did not have electronic attendance records captured in the SMIS and had to be removed. The attendance data also varied widely from student to student and may have contributed towards reducing the accuracy of the models.

- For some categories of degree classes of interest to the prediction problem, there were not enough cases to train the model. For instance, the number of first class degrees out of the total population was just 62, the result of this is that sometimes when test data was sampled from the total population, no first class degree students were sampled

- ELearning portal access data would also have been a good attribute to test for its impact on student performance. Since eLearning portal access data was not available, this could not be possible.

5.3 Suggestions for Improvement

There may be cases where an institution wishes to implement the solution proposed here but does not have all or some of the data required for data mining. Such institutions can overcome this challenge by use of an alumni programme through which they can collect some or all or the data about former graduates. This data is then used to update the student data warehouse.

The factors that affect student performance may vary between different campuses of the same institution; in fact they may vary even between different degree programs within the same campus of an institution. On this basis, perhaps the results of the data mining process can be further improved by the creation of campus specific or program specific data marts. In this same vein, another distinction that can be made is to differentiate between regular and evening program students then mine them separately.

5.4 Recommendations and further research

Mining to predict student overall performance in terms of the final degree classification they are likely to get is just but a starting point. This can be taken further to the level of predicting student performance in a specific year, semester or even predict the performance in a particular course. Since the problem is still one of prediction, the methodology employed in this study can still be employed with some minor changes. The resulting prediction models for all the different levels final degree class, yearly, semesterly and course can then all be integrated into a single solution. By so doing, the tutors will be armed with far much more knowledge to use in student academic advising to hopefully improve student performance.

With further developments in the field of EDM, the entire process of student academic advising can be done by a computer program that makes use of data mining models, a knowledge repository and a data warehouse to generate an advice report with little or no human intervention.

APPENDICES

I Project Plan

Shown below is a schedule that will guide the project:-

Year	**2013**																							
Month	**February**				**March**				**April**				**May**				**June**				**July**			
Week	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Project Proposal																								
Development of the Data Warehouse																								
Extraction, Transformation and Loading of Data																								
Testing and Evaluation																								
Data Mining																								
Interpretation and Reporting of Results																								
Incorporation into Learning Management System																								
Project Report Completion and Submission																								

Table 16: Project Plan

II Project Budget

The table below shows the financial requirements for implementing the project:-

ITEM	**COST**
Server Computer	70,000/=
Books and References	10,000/=
Miscellaneous	10,000/=
Document Production	15,000/=
TOTAL	**105,000/=**

Table 17: Project Budget

REFERENCES

1. Aher, S. B., & Lobo, L. (2011). *Data Mining in Educational System using WEKA.* Paper presented at the IJCA Proceedings on International Conference on Emerging Technology Trends (ICETT)(3).

2. Badur, B., & MARDIKYAN, S. (2011). *Analyzing Teaching Performance of Instructors Using Data Mining Techniques.* Informatics in Education-An International Journal (Vol 10_2), 245.

3. Baepler, P., & Murdoch, C. J. (2010). *Academic analytics and data mining in higher education.*

4. Baker, R. (2010). *Data mining for education.* International Encyclopedia of Education, 7, 112-118.

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

6. Baradwaj, B. K., & Pal, S. (2012). *Mining Educational Data to Analyze Students' Performance.* International Journal of Advanced Computer Science, Vol. 2, No. 6

7. Barracosa, J., & Antunes, C. (2011). *Anticipating Teachers' Performance.* Department of Computer Science and Engineering, Technical University of Lisbon, Portugal.

8. Chandra, E., & Nandhini, K. (2010). *Knowledge Mining from Student Data.* European Journal of Scientific Research, 47(1), 156-163.

9. Ismail, R. J. (2010) *Mining Tutors' Interesting Areas to Develop Researched Papers Using a Proposed Educational Data Mining System.* Engineering and Technology Journal, Vol. 30, No. 10

10. Lee, C.-H., Lee, G.-G., & Leu, Y. (2009). *Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning.* Expert Systems with Applications, 36(2), 1675-1684.

11. Nguyễn, T. B., Hoàng, T. Á. D., Trần, H., Nguyễn, Đ. N., & Nguyễn, H. S. (2008). *An integrated approach for an academic advising system in adaptive credit-based learning environment.* VNU Journal of Science, Natural Sciences and Technology, Vol. 24, pp. 110-121

12. Pandey, U. K., & Pal, S. (2011). Data Mining: *A prediction of performer or underperformer using classification.* International Journal of Computer Science and Information Technologies

13. Ramaswami, M., & Bhaskaran, R. (2009). *A study on feature selection techniques in educational data mining.* Journal of Computing, Vol. 1, Issue 1

14. Ramaswami, M., & Bhaskaran, R. (2010). *A CHAID based performance prediction model in educational data mining.* International Journal of Computer Science Issues,

15. Ranjan, J., & Khalil, S. (2008). *Conceptual framework of data mining process in management education in India: an institutional perspective.* Information Technology Journal, 7(1), 16-23.

16. Romero, C., & Ventura, S. (2007). *Educational data mining: A survey from 1995 to 2005.* Expert Systems with Applications, 33(1), 135-146.

17. Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40*(6), 601-618.

18. Romero, C., Ventura, S., & García, E. (2008). *Data mining in course management systems: Moodle case study and tutorial.* Computers & Education, 51(1), 368-384.

19. Su, J.-M., Tseng, S.-S., Wang, W., Weng, J.-F., Yang, J. T. D., & Tsai, W. (2006). *Learning portfolio analysis and mining for SCORM compliant environment.* Journal of Educational Technologyand Society, 9(1), 262.

20. Superby, J., Vandamme, J., & Meskens, N. (2006). *Determination of factors influencing the achievement of the first-year university students using data mining methods.* Paper presented at the Workshop on Educational Data Mining.

21. Tair, M. M. A., & El-Halees, A. M. (2012). Mining Educational Data to Improve Students' Performance: A Case Study. *International Journal of Information, 2*(2).

22. Talavera, L., & Gaudioso, E. (2004). *Mining student data to characterize similar behavior groups in unstructured collaboration spaces.* Paper presented at the Proceedings of the Artificial Intelligence in Computer Supported Collaborative Learning Workshop at the ECAI 2004.

23. Tang, C., Lau, R. W., Li, Q., Yin, H., Li, T., & Kilis, D. (2000). *Personalized courseware construction based on web data mining.* Paper presented at the Web Information Systems Engineering, 2000. Proceedings of the First International Conference on.

24. Tanimoto, S. L. (2007). *Improving the prospects for educational data mining.* Paper presented at the Data Mining for User Modeling On-line Proceedings of Workshop held at the.

25. Ueno, M. (2004). *Data mining and text mining technologies for collaborative learning in an ILMS ''samurai''.* Paper presented at the Proceedings of the IEEE International Conference on Advanced Learning Technologies.

26. Vialardi, C., Bravo, J., Shafti, L., & Ortigosa, A. (2009). *Recommendation in higher education using data mining techniques.* Paper presented at the Proceedings of the 2nd International Conference on Educational Data Mining (EDM09).