CERTIFICATION OF APPROVAL

Data Mining and Prediction Tools
(for Predicting Students' Success in Programming Course)

By

Che Sarah Che Nordin

A project dissertation submitted to the

Business Information System Programme

Universiti Teknologi PETRONAS

in partial fulfilment of the requirement for the

BACHELOR OF TECHNOLOGY (Hons)

(BUSINESS INFORMATION SYSTEM)

Approved by,

(Ms Elaine Chen Yoke Yie)

UNIVERSITI TEKNOLOGI PETRONAS

TRONOH, PERAK

May 2011

Abstract

This project addresses the importance of extraction and analysis of data from different types of educational settings such as computer-based or web-based educational system (i.e. course management system), classroom environment factors as well as psychosocial factors in the university, which can affect the students and use these data to foresee students' learning patterns.

The vast amount of data from different educational settings can be fully utilized to predict the students' performance or a particular course. However, there is no tool as such, that can automatically manage, extract and analyze this kind of information. Besides that, most of the current data mining tools are too complex for educators to use and their features go well beyond the scope of what an educator might require. This project will use the data mining approach and techniques in analyzing different types of data gathered from different educational settings.

The project aims to develop a new data mining and prediction tools, which will analyze different types of data coming from different educational settings to assist lecturer to predict students' performance in a programming course.

The scope of study for this project is one of the programming courses in the university, Advanced Business Application Programming (ABAP) and the university's E-Learning System.

The main contribution of this project is the development of a new data mining and analysis tools, that can produce prediction output to assist the lecturer in his or her decision making activities to improve the learning process in a particular programming course.

Acknowledgement

Completing this Final Year Project has been one of the significant achievements during my four years of study at Universiti Teknologi PETRONAS. This will never be possible without the help of some of the most influential people in my life and of course the Most Almighty God because with His Blessing and Help, I have managed to finish the project as well as completing this dissertation.

I owe my deepest gratitude to my supervisor, Ms. Elaine Chen Yoke Yie, whose encouragement and guidance from the initial to the final phase of this project enabled me to develop a good level of understanding on project management and processes. She has made available her support in a number of ways as well as improving my ability to work better independently as an individual and in a team. I am very grateful to be presented with this learning opportunity and it is an honor for me to be given such trust from her and in return, give some contributions by completing this project. I am also indebted to others who have helped me along the way, especially to Ms. Syakirah Mohd Taib, which is my Data Mining lecturer in giving me guidance as well as advises throughout the process of completing this project. Besides that, I also would like to thank all the examiners whom have given me comments and constructive criticism in order to improve the system.

I also want to express my appreciation to my beloved family for the never-ending support and help that they have given me throughout my whole years of study and making this project an accomplished mission.

Besides that, I would like to thank the coordinators for Final Year Project 1 and 2 as well as Universiti Teknologi PETRONAS for providing us with all the important information and guidelines and thanks to everyone who has contributed directly or indirectly in completing this task.

TABLE OF CONTENTS

Table of Figures

Table of Tables

CHAPTER 1

INTRODUCTION

1.1 Background

According to [1], Educational Data Mining is an emerging discipline, concerned with developing methods for exploring unique types of data that come from educational settings, and using those methods to better understand students, and the setting which they learn in. Whether educational data is taken from students' use of interactive learning environments, computer-supported collaborative learning, or administrative data from schools and universities, it often has multiple levels of meaningful hierarchy, which often need to be determined by properties in the data itself, rather than in advance. Issues of time, sequence, and context also play important roles in the study of educational data.

Course management systems (CMSs) can offer a great variety of channels and workspaces to facilitate information sharing and communication among participants in a course. They normally also provide a database that stores all the system's information: personal information about the users (profile), academic results and users' interaction data. However, due to the vast quantities of data these systems can generate daily, it is very difficult to manage manually [2].

Educational Data Mining researchers study a variety of areas, including individual learning from educational software, computer supported collaborative learning, computer-adaptive testing and the factors that are associated with student failure or non-retention in courses. Across these domains, one key area has been in the improvement of student models. Student models represent information about a student's characteristics or state, such as students' current knowledge, motivation, meta-cognition and attitudes. Modelling student individual difference in these areas enables software to respond to those individual differences, significantly improving student learning [17].

1.2 Problem Statement

Different types of educational settings, such as the traditional classroom environment or CMS can provide a vast amount of useful information that are valuable and can be fully-utilized in analyzing students' learning patterns in their learning process. However, there is no tool as such, that can automatically manage, extract and analyze this kind of information. The implementation of data mining in this area is useful in assisting the lecturer to foresee students' learning patterns and adjust the teaching methods to suit different learning styles in a particular group of students.

Instructors and course authors demand tools to assist them in this task, preferably on a continual basis. Nowadays, data mining tools are normally designed more for power and flexibility than for simplicity. Most of the current data mining tools are too complex for educators to use and their features go well beyond the scope of what an educator might require. As a result, the CMS administrator is more likely to apply data mining techniques in order to produce reports for instructors who then use these reports to make decisions about how to improve the student's learning and the online courses [2].

1.3 Proposed System

The importance of this project is to addressed the problems mentioned above by developing a new data mining and prediction tools, which will analyze different types of data coming from different educational settings, such s from traditional classroom environment, student psychosocial factors as well as data extracted from online learning environment or CMS, and produce a prediction output on student's final grade that can be used to foresee students' learning and provide timely intervention to adjust the teaching methods and learning process and thus, achieving the goals in a particular programming course.

1.4 Objectives

The project aims to develop a new data mining and prediction tools, which will analyze different types of data coming from different educational settings to assist lecturer to predict student's performance in a programming course.

1.5 Project Relevancy

Educational data mining is an emerging discipline, concerned with developing methods for exploring unique types of that come from the educational context [2]. Most of the systems that exist in the current market do not integrate data mining tools in order for the user to capture useful information that can be used for educational purpose. Hence, there is a demand for the data mining tool to be developed to serve this purpose. The development of such system can give beneficial contribution to the field of educational data mining as a whole by producing information that can be used to improve teaching and learning process by the lecturers to help students to succeed in a particular course.

1.6 Project Feasibility

When deciding to build a new analysis and data mining tool, it is critical to have good insights on the potential end users. This will most likely help to create an attractive and appealing system for the intended users. Hence, in the case of our proposed system, the development of a new data mining and prediction tools will include the evaluative resource which draw upon statistically relevant CMS tracking data, classroom environment data as well as psychosocial variables data, to facilitate more meaningful real-time pedagogical analysis that will allow the lecturer to monitor student engagement and learning progression, as well as to evaluate the impact of implemented learning and teaching activities [3], specifically for lecturer teaching programming course.

Scope of study, Context and Data Source

The scope of study in this project will include the stakeholders in one of the programming course in UTP, which is Advanced Business Application Programming (ABAP). The data will be gathered around the individuals involved in this class, which are the lecturer as well as the undergraduate students. This subject is one of the major subjects in E-Business major for Business Information Systems programme. The current enrolment of students in this course is approximately around 30 students. The course is delivered in face-to-face classroom environment setting with the supplementary of UTP e-learning system. All the data for this project will be retrieved from this course as well as UTP e-learning system which is using Moodle software package.

Time Feasibility

The project is expected to due somewhere in August 2011. Hence, the duration of the project will be around five months starting from the project initiation on the 2nd week of February 2011. The project is divided into different phases which are, planning phase, analysis phase, design phase and implementation phase. A Gantt chart on the project schedule is crafted (see Appendix 1). The Gantt chart will be revised from time to time ensure that it is able to meet the requirements of Final Year Project 1 and Final Year Project 2 deliverables submissions.

CHAPTER 2

LITERATURE REVIEW

Traditional methods of evaluating and predicting students' performance

Even before the emergence of technological way to support learning process, such as CMS, the educational system has been using a few methods in monitoring and predicting student success. In a university, student's overall performance is determined by internal assessment as well as external assessment. Internal assessment is made on the bases of a student's assignment marks, class quiz, attendance, previous semester grade and his/her involvement in extra curriculum activities. While at the same time external assessment of a student based on marks scored in final exam is also taken into calculation. This correlated information should be conveyed to the class teacher before the conduction of final examination. This study often helps the teachers to reduce the failure rate and improve the performance of the students [20].

The model identifies the weak students before final exam in order to save them from serious harm. Teachers can take appropriate steps at right time to improve the performance of student in final exam. It deals with both kind of assessments especially internal assessment in order to predict students whose performance is low. This model check the performance of student at different levels before final exam in order to predict weak students and take appropriate steps to save them from failure [20].

Technological methods of evaluating and predicting students' performance

With the technological advancement, the newest data management tools allow information sources including ones that have not traditionally been viewed as achievement-related, to be accessed at the same time, facilitating multidimensional analysis. The most obvious example of this is the integration of achievement data

with demographic information from a school's students information system, making it possible to sort results by gender, race and much more [16].

The changing factors in contemporary education has led to the quest to effectively monitor student performance in educational institutions, which is now moving away from the traditional measurement and evaluation techniques to the use of Data Mining Techniques which employs various intrusive data penetration and investigation methods to isolate vital implicit or hidden information [23].

The main attribute of data mining is that it subsumes Knowledge Discovery which is a nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data processes, thereby contributing to predicting trends of outcomes by profiling performance attributes that supports effective decision making. Performance profiling is dependent upon motivation, attitudes, peer influence, curriculum and by the response system and predicts correctly which students may need some attention or reinforcements in the course of their education [23].

The process of tracking and mining and mining student data in order to enhance teaching and learning is relatively recent but there are already a number of studies trying to do so and researches are starting to merge their ideas. A timely and appropriate warning to students at risk could help preventing failing in the final exam. Therefore it seems that data mining has a lot of potential and benefits for education [21].

A predominant invention in the area of Educational Data Mining is Course Management System (CMS). CMS is a system that holds vast quantities of data on a particular course in web-based or computer-based setting such as, data on server logs, total time online, total discussion messages posted, number of assignments submitted, grades and many more. It is an online environment that allows information to be shared between students and lecturer [5]. CMS has often been used as a medium to support the traditional classroom learning environment or to conduct distant learning courses.

According to [5], some possibilities that this system can accomplish in order to support teaching and learning are transmission of information, evaluation of teacher and learner performance, and interaction online. The most widely used and purpose of using this system is for transmission of the course materials such as lecture notes, lecture slides and other course handouts to the students. Some examples of commercial systems are Blackboard, WebCT, Top-Class, while some examples of free systems are Moodle, Ilias and Claroline [2].

Comparison of works – similarities and differences of existing systems

Three of the most commonly used CMSs are Blackboard or WebCT (now owned by Blackboard) and Moodle. The similarities and differences of these systems are discussed further in this section.

Some of the similarities on its features are such as uploading files (e.g. Word documents, PowerPoint, audio file), linking to external web sites, discussion forum, synchronous chat, quizzes and tests, drop box, course calendar, grades, monitoring student participation, copy course over from semester and customized template.

The differences between Moodle and WebCT systems are explained in the table below [13]:

Feature	**Moodle**	**WebCT**
Bandwidth	All features work on dial-up	Bandwidth hog: slow to load, times out
Learning curve	Can use without manual or training; no configuration	Not as intuitive because many components; need to configure initially
Discussion	Photos, nested threads	See posts one at a time; not nested
Tools	Blog, wiki, journal, glossary, workshop (Wimba coming soon), RSS feed	Whiteboard & hands raised in chat; Wimba for audience discussion; one-on-one in group

Cost	Free	Need to pay for the license fee
Customization	Open source (so can change locally, (Twin Cities) on demand by IT programmer)	Need to request change from WebCT (may not happen or happen quickly)
User Statistics	Chart comparing students; no. of visits per page	Time student spent on each page
Quizzes	Quizzes & tests; Vote option; tests built into lesson	Quizzed & tests; can build in tests using other tools

Table 1: Comparison between Moodle and WebCT

Due to the fact that these systems can generate numerous amounts of data, the user can make full use of this system to improve the learning process. [3] confirms the earlier studies that have been done on predictive power of learning management system (or course management system) data to develop reporting tools that can identify at-risk students and allow more timely intervention on learning methods from lecturer. Although some platforms offer some reporting tools, it becomes hard for a lecturer to extract useful information when there are a great number of students. They do not provide specific tools allowing lecturer to thoroughly track and assess all students' activities while evaluating the structure and contents of the course and its effectiveness for the learning process. A very promising area for attaining this objective is the use of data mining [2]. According to [2], this method helped to uncover new, interesting and useful knowledge based on students' usage data. Some of the e-learning problems or subject to which data mining techniques have been applied are dealing with the assessment of students' learning performance, provide course adaptation and learning recommendations based on the students' learning patterns, dealing with the evaluation of learning materials and educational web-based course, provide feedback to both lecturers and students of e-learning courses, and detection of a typical students' learning pattern.

Data mining in CMS is the process of extracting data from course management system, such as e-learning and analyzes these data to find different patterns and relationships between each data.

Critical analysis of literature

Despite all the benefits in the use of CMS mentioned in the literature, there were also problems and limitations with regard to the studies that have been done in this educational data mining area. According to the study done by [3], the limitations gave impact in the overall generalizability and interpretation of the findings. For example, the implications of the study are limited by its focus on data derived from a fully online course within one institution. In fully online course, it is reasonable to expect that the only venue for student interaction with peers, lecturers and course content is via the CMS. Future studies should be directed towards the investigation and analysis of potential significant CMS tracking indicators measures in relation to student success for alternate pedagogical designs and course modalities. On top of that, another problem discussed in [4] is that in the current CMS, lecturer would be able to get quick view of basic learning data such as login date, frequency of visits and etc. However, there is no function or feature that is available to help instructors in identifying learners' individual or group learning patterns or to identify successful or less-successful and identify necessary facilitation needed.

According to [2], data mining tools are normally designed more for power and flexibility than for simplicity. Most of the current data mining tools are too complex for lecturer to use and their features go well beyond the scope of what a lecturer might require. As a result, the CMS administrator is more likely to apply data mining technique in order to produce reports for lecturers who then use these reports to make decisions about how to improve students' e-learning and the online courses.

The studies done by [2], [3], [4], and [5] were all using separated data mining process to analyze the data available in CMS. Hence, the researchers in [4] strongly suggest that CMS developers should integrate data mining tools to facilitate effective online teaching and learning.

Based on the limitation of the study done by the [3], in order to increase the predictive power of CMS, the future development of CMS should include data coming from other educational setting such as classroom environment factors as well as psychosocial factors that also affect the students' learning process.

As mentioned earlier in this proposal, one of the goals of the system is to foresee students' learning pattern as well as predicting students' final grade.

A student success in this project context, which is a student's success in a particular course, is measured by the final grade attained by the student at the end of the course (course completion). Besides analyzing the factors or data that are extracted from CMS, we need to also consider some other factors such as data coming from classroom environment as well as the psychosocial factors. Psychosocial refers to one's psychological development in and interaction with a social environment [12]. As we shall see further, all these data are indeed interrelated with one another.

Based on the research done by [10], it indicates that data from classroom environment such as student attendance is statistically significant in explaining class grade and overall performance of students. Students who missed class frequently significantly increase their odds of a poor grade in a particular course. In [10], it is stated that the most valuable and important determinant of student success is each unit of time spent in the class itself, followed by any time spent in discussions sections that accompanied the lectures, the time spent outside of class preparing for the class itself and the least significant time commitment in improving student performance in a particular class was the time spent studying for the final exam. It was concluded in [10], that the most important learning in a course takes place in the classroom and that students who do a conscientious job on a daily basis preparing and participating in class outperform those students who skip class and try to cram for examinations. This is further supported by [9], in which with attendance having major influence on academic performance, even to the point of some professors using it as a requirement to pass a course.

The results of studies obtained on how psychosocial factors affect students learning pattern and final grade were ambiguous and inconsistent. The study done by [6] was also unable to identify the relationship between personal characteristics and final course grade. However, these data should be taken into calculation in order to increase the accuracy of foreseeing students' learning pattern as well as predicting

students' final grade. As stated in [11], identifying these factors has the potential to be useful in several important ways. First, it can provide a basis for helping students to reflect on their perceptions and expectations of university study so that they can gain more control over their learning and approach university's studies in a way that will maximize their chances of success. Second, it can provide a basis for helping lecturers to reflect on their expectations of and about students so that they will be better informed about ways in which they can facilitate student learning, enhance the influence of positive factors and minimise the influence of negative factors on student success. Third, the results can be used by university administrators to help them to provide a learning environment that will maximise the chances that students will be successful.

Based on the study done by [7], the factors that instructors believe can influence students' success fell into five categories which include:

- The nature of the subject
- Intrinsic characteristics of the student
- Student background (or previous experience)
- Student attributes and behaviours
- Developmental strategies used by the instructors to help students succeed

The nature of the subject being taught is the characteristics of specific topics or concepts, such as algorithm analysis, pointers and concurrency, which can present unique challenges or obstacles to students' success.

Intrinsic characteristics of the student is the personality traits such as being positive, inquisitive and motivated are some of the factors that can contribute to success. Besides that, culture and language are also thought as the factors influencing students' comfort level and willingness to ask questions. This is further supported by [19], which examined on the importance of two cognitive factors, motivation and comfort-level in a student's learning process. Being proactive and asking questions

in class suggests that students are more comfortable with the class environment. Study in [8] indicated that students who are actively engage in the learning process were observed to have positive correlation with their final CGPA. On top of that, many instructors mentioned that there is a special ability in some students who can capture knowledge so easily while there are some students, whom after many ways and efforts in studying the subject are still unable to really understand the whole concept. However, studies investigating gender differences, age and years in college seem to have no significant relevance for predicting success [7].

Student background includes the student previous experience that can give impact to their success. Instructors in this study [7] acknowledged that even just a bit of programming background can be helpful. In spite of that, the finding by [6] indicated that student prior experience in programming before university can contributes to a good grade but not necessarily since some students who do not have programming experience can excel in programming subject as well. Besides that, student background factor such as previous school attended can also depict the learning pattern of a student. Students who attended boarding school might have different learning style compared to students who attended the normal daily secondary school. But this is subject to the students' ability to adapt to changes and being independent. The author in [18] supported this by mentioning that prior academic experience and prior computer experience may be the factors that influence the performance of student in programming course.

Student's attitudes and behaviours are also considered as important determinants to succeed in studies. Hard work and persistence are very important as good course work marks (which largely contribute to the course final grade) is attributed from hard work and were found to be predictors of success in a programming course. Interest in and general positive attitudes toward programming are also important for success. Likewise, the positive expectation of an 'A' in the course was found to be a contributor to success for some students. On the other hand, negative emotions such as lack of interest in programming, and afraid to try or learn new things using computer were mentioned as hindering success [7].

Many instructors naturally considered the influence of their own actions such as how they might explain things more clearly, employ appropriate pedagogical techniques to different learning styles or assessing students' progress [7].

On top of that, other unidentified factors can also contributes to student achievements in programming course such as learning environment, motivation, learning facilities and instructor ability which need to be considered by the faculty [6]. Study in [9] also stated other factors that can affect students' learning process such time management, financial problems, sleep deprivation, social activities as well as health-related factors that include amount of exercise, nutritional routines and amount of social support can affect students' learning process.

Summary

The plan to develop a new data mining and prediction tools that can provide the lecturer with useful information that can be used in their decision making process is one of the unique factors that inspire me to move to the design phase of this project. Besides that, the system will also have a capability to take into consideration and analyze many different types of data coming from different educational settings that can further affect the students' learning process. The identification of such factors makes it possible to develop a tool to provide an early diagnosis of a student's likely performance on a programming module. Interested parties could use the tool to make more informed decisions on appropriate course of actions and to decide upon personalized interventions that foster a student's intellectual strengths [18].

Since majority of the systems available in the current market do not have data mining tools built in together in it and the fact that the separate data mining tools that are available in the market are considered by most users as too complex to be used, the development of this new tools that can analyze different types of data (taking into consideration the different types of factors that can affect the learning process) is seen as a great benefit for the lecturers.

CHAPTER 3

METHODOLOGY

This project adopts hybrid Rapid Application Development (RAD) system development life cycles which include Prototyping-based and Throwaway Prototyping-based methodologies. Phase 1 mainly involves research work while phase 2 involves the data collection and analysis, development of analysis models and deployment of data mining techniques to all the data gathered in phase 1. Phase 3 and 4 make up the main development (design and implementation) stage.

The project's methodology is based on a relatively thorough analysis phase that is used to gather information and to develop ideas for the system concept (as in the case of throwaway prototyping methodology). The issues in data mining, clarifying user requirements, understanding the technology or programming language to be used and analyzing the complexity of the system will be analyzed and solved in this phase. The information based on the thorough analysis and design phase will be used to produce the system prototype, which will be further refined until the prototype provide enough functionality to be considered as a complete system. The main reason of combining two methodologies in this project is because the design prototypes will not be thrown away (as in the case of throwaway prototyping), and instead it will evolve into the final system (as in the case of prototyping methodology) [15].

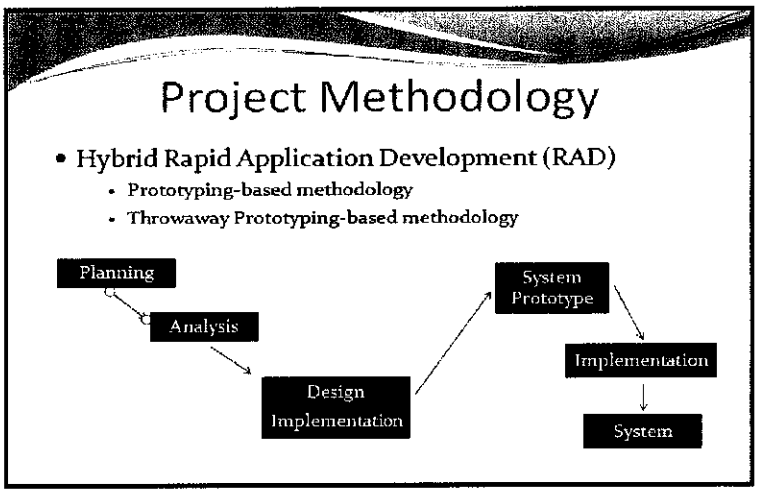


Figure 1: Project Methodology

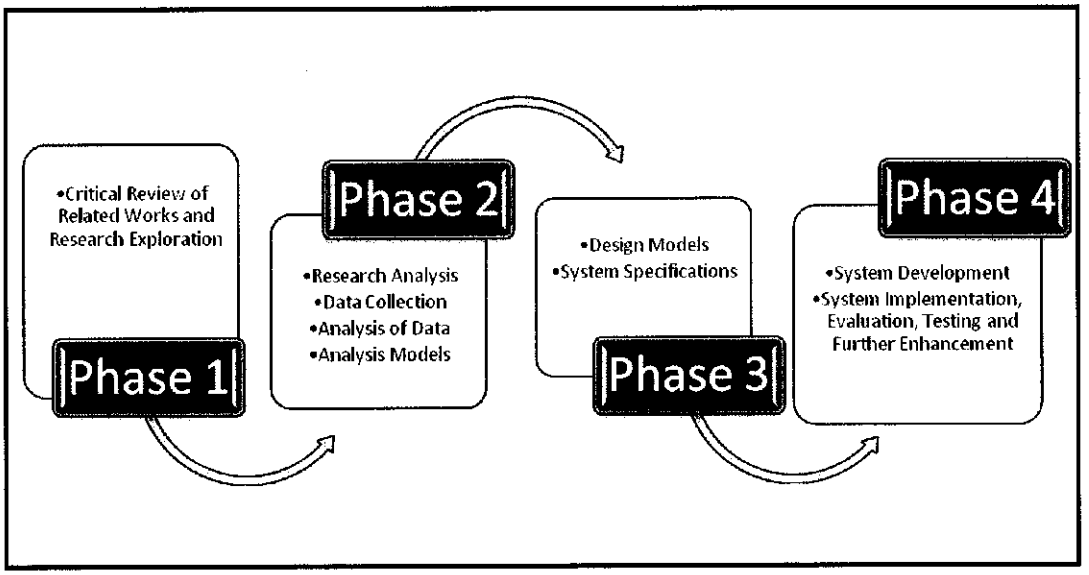


Figure 2: Project Activities

Phase 1: Critical Review of Related Works and Research Exploration

The project is initiated with a detailed background study on the features of existing CMS available in the market, possible tracking data that can be extracted from CMS, possibilities of how these data can help in improving students learning process, data mining techniques applied in the area of educational data mining, study on the factors that can predict student success and learning patterns, analyzing and

comparing the studies in this project with other related works in the same area, as well as discovering gaps in the existing literature.

Phase 2: Research Analysis, Data Collection, Analysis of Data and Analysis Models

This phase includes the process such as developing research strategy (purpose, time frame, scope, and environment), data collection design (i.e. interview, questionnaires, analysis on the existing system), data collection and preparation, data analysis and interpretation (i.e. the deployment of data mining techniques), and the development of analysis models. The analysis models describe how the new system will be developed with the use of functional modelling, structural modelling as well as behavioural modelling methods [15].

Figure 2 and 3 on the next pages portray the basic functions of the current (as-is) system – that is, what the users can do and how the system should respond to the users' actions. These two diagrams represent how a system interacts with its environment.

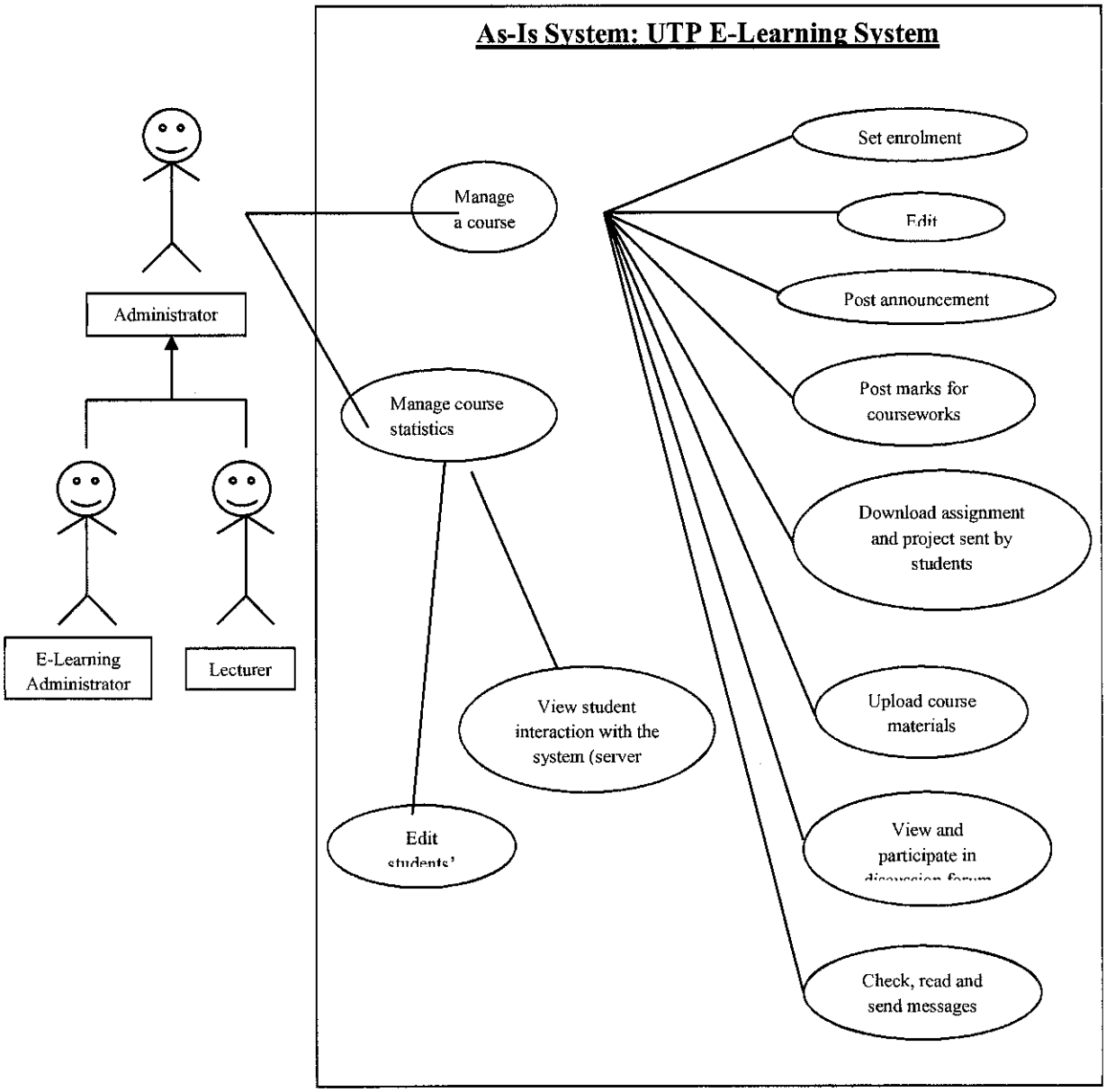


Figure 3: As-Is System (Part 1)

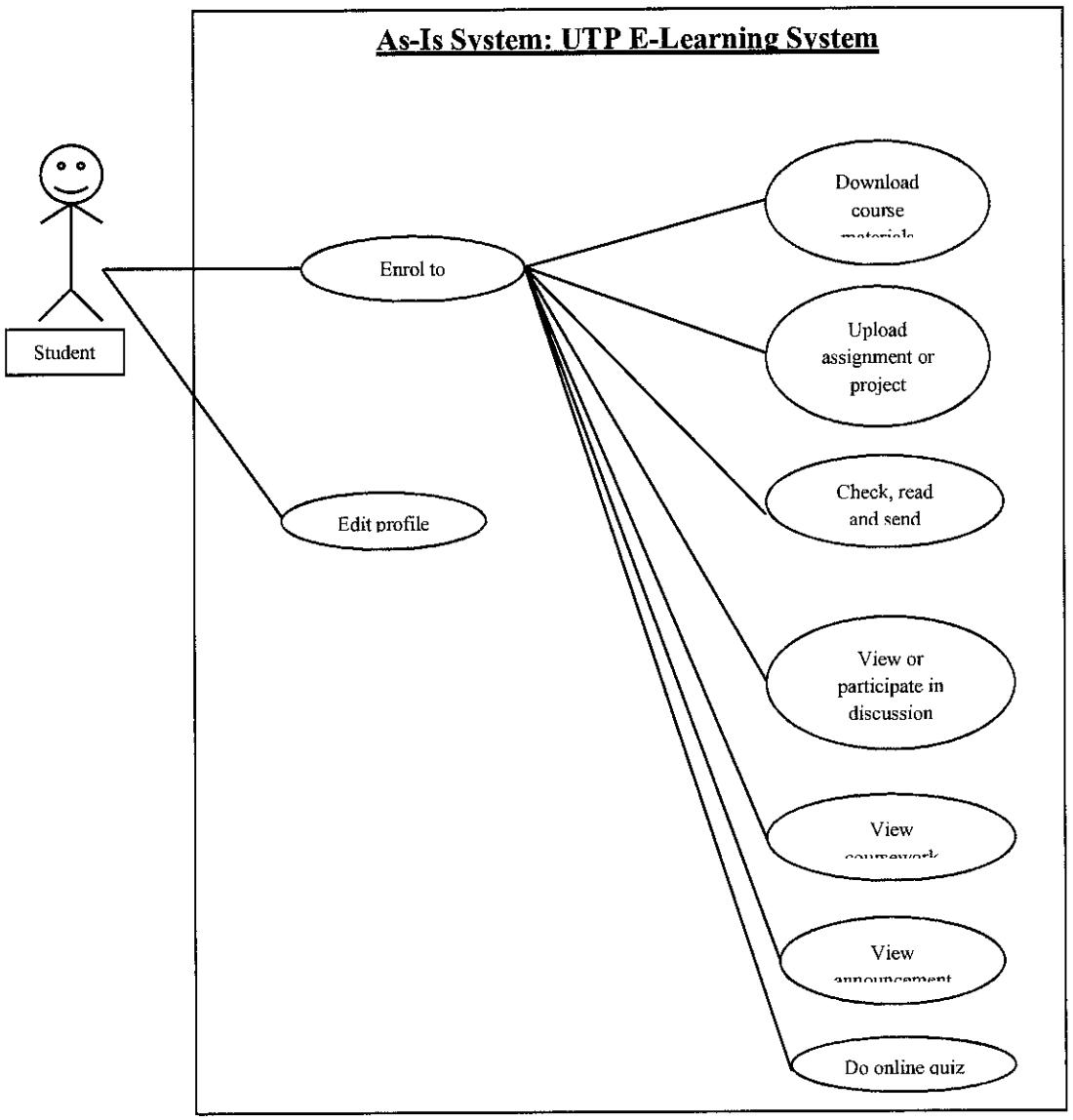


Figure 4: As-Is System (Part 2)

As a result of the interview session with UTP E-Learning Administrator (see Appendix 2) some useful information were gained about the features of the system (analysis of the current or existing system). As depicted in Figure 2 and 3, the system can be accessed by two types of users which are the administrator as well as the student. The administrator can be further divided into two subcategories which are the system administrator and the course lecturer. The person who holds the highest level of access is the system administrator, which has the authority to manage and

oversee all the access into all courses available in the system. The access of the lecturer is limited to the course taught only. Besides that, other users of the system are the students of the university.

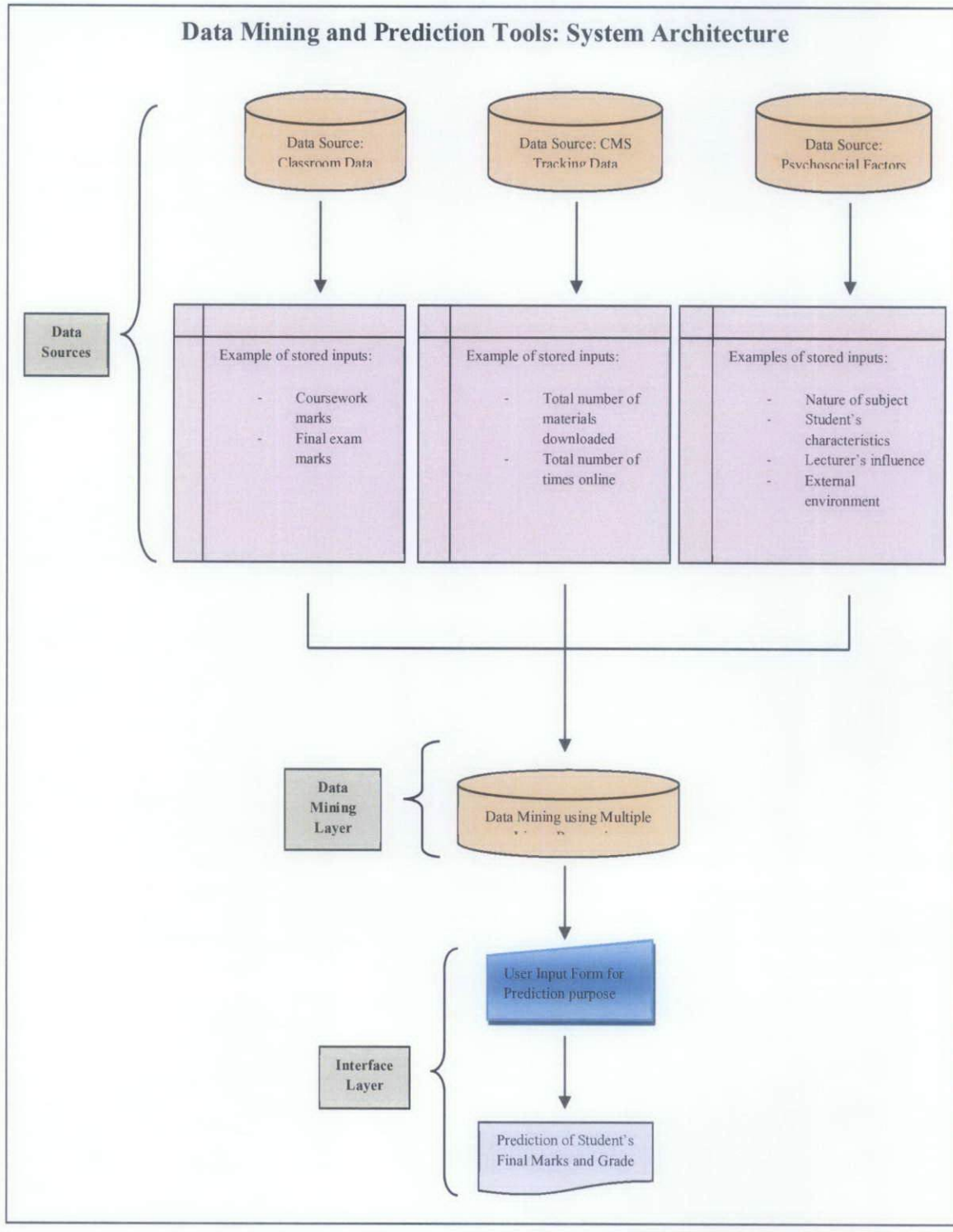


Figure 5: System Architecture

Figure 5 portrays the system architecture of the proposed system. As depicted in the figure, the data source will be collected from class session, CMS tracking data as well as psychosocial factors. The data from the class session such as records coursework marks (which include students' quizzes, tests, assignments, lab test and projects marks) and final exam marks will be obtained from the lecturer of ABAP course. The CMS tracking data will be obtained from the CMS in the university, which is UTP E-Learning System. Data from the server logs were studied and two variables which seen as relevant to be used in this project were extracted. The two variables are total number of materials downloaded and total number of times online. Data of students' psychosocial factors are also found to be one of the important determinants in predicting the students' performance for the course. These data are such as nature of subject, student's characteristics, lecturer's influence and external environment. These data were captured using one of the common research techniques, the questionnaire.

The data mining layer is the layers whereby the data analysis and interpretation will be done. The deployment of data mining technique, multiple linear regression will occur in this layer.

The interface layer will enable the user, which is the lecturer to input some data in order to calculate prediction for student's final marks as well as a page to display the prediction output.

Phase 3: Design Models and System Specifications

In this phase, the data gathered from the phase 1 and 2 will be used in the modelling and designing the whole architecture of the system. The design phase decides how the system will operate, in terms of user interface, forms and the program design which defines the programs that need to be written and exactly what each program will do [15].

Phase 4: System Development, System Implementation, Evaluation, Testing and Further Enhancement

System construction is the first step in implementation phase. The system is built and tested to ensure it performs as designed. Testing is one of the most critical steps in implementation to remove any bug in the program. Support plan may also be included in this phase. The plan usually includes a formal or informal post-implementation review, as well as a systematic way for identifying major and minor changes needed for the system [15].

Tools or equipment required

This project is developed and implemented using HTML, javascript as well as CSS coding for the design part. The javascript functions include all the necessary functions to perform data mining using multiple linear regression and to build functions to calculate students' final marks and classify the marks based on UTP grading schemes.

Theoretical foundation and hypotheses development

The vast amount of data available in different kind of educational settings namely, the classroom environment, online learning environment (CMS tracking data) as well as the psychosocial factors can be used in order to predict the students' performance and their final grade or marks for a particular course. Based on the literature review done, most of the research papers have done a number of studies on how factors in the classroom environment and online learning environment might have an impact in predicting the student's final grade in a particular course. However, psychosocial factors should also be taken into consideration in the prediction of student's final grade. Psychosocial factors relates closely to one's psychological development in and interaction with his or her social environment. Psychosocial factors can be the student's attributes and behaviours, student's motivation, student's background,

student's intrinsic characteristics as well as the nature of the subject. With this in mind, a framework that provides a foundation in the prediction of student's final grade is proposed. The theoretical framework is shown in the figure below. The framework also shows the hypotheses paths of the attributes.

Theoretical model was developed to test how different data from different educational settings affect the prediction of student's performance. The attributes tested are the coursework marks, questionnaire result as well as the CMS tracking data. The research model for this study is depicted in the figure below:

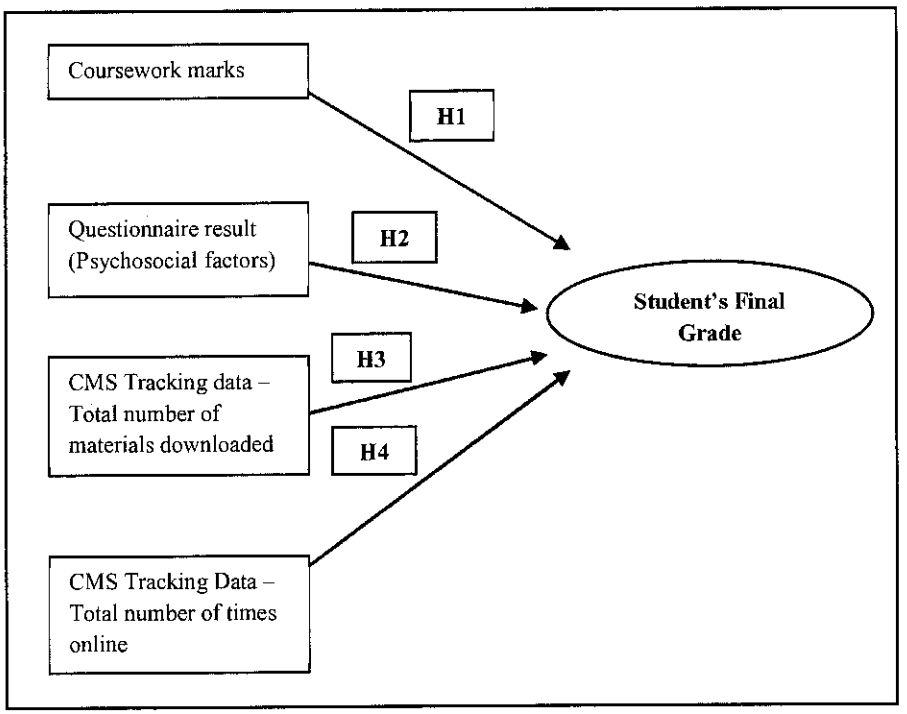


Figure 6: Research Model

Coursework marks: Coursework marks is a major assessment criteria in a particular course for this university. Coursework marks carry a weightage of 50% from the overall marks of 100%. For ABAP course, the data included in calculating the total coursework marks are quizzes marks, tests marks, assignments marks as well as the lab test marks. Therefore, the coursework marks is seen as a significant indicator on how a student will perform in the final examination and can be used as a strong factor in predicting the student's final grade. A student needs to have a strong

coursework marks in order to obtain a good grade in a particular course. Therefore, this hypothesis is proposed:

- **H1:** Coursework marks have a direct correlation in the prediction of a student's final marks or grade.

Questionnaire Result: The questionnaire was designed in a manner that it covers different areas of psychosocial factors that might have an impact in predicting the student's performance. There are basically four main sections in the questionnaire. The first section covers the factors on the nature of the subject (Question 1). The second section covers the factors on student's characteristics, attributes and behaviours (Question 2, 3, 4 and 5). The third section covers the factors on lecturer's influence towards the student's learning process (Question 6, 7 and 8). And the last section covers the factors on external environments (Question 9, 10, 11 and 12).

All of the four sections above were seen to have a probability in determining the student's performance on a particular course. Students needs to have high confidence that they are able to grasp the whole concepts in that subject (nature of the subject), positive characteristics, attributes and behaviours towards learning programming and learning the subject specifically, have confidence about the lecturer's influence in helping them to succeed in a particular programming course and are equally comfortable or satisfied with the external environment in which they believe might affect their learning process. Therefore this hypothesis is developed:

- **H2:** Questionnaire result has a direct correlation in the prediction of a student's final marks or grade.

CMS Tracking Data (From UTP E-Learning System): Previous studies have shown that CMS tracking data have a predictive power in determining student's success in a particular course. From studying the data available in CMS, we can see the student's activities throughout the whole semester with regard to online learning environment. For ABAP course, CMS served as a supplementary medium in order to distribute the course materials as well as the course updates to the students. The students used the system to prepare for the class as well as a part of their preparation

before taking tests, quizzes and etc (e.g. download materials for reference and revision). The students that actively used the system are seen to have a better preparation in their studies (e.g. Read notes before entering the class, read notes consistently throughout the whole semesters). Better preparation should always lead to better results in the final examination.

Although there are a lot of data that can be extracted from the system, two data are seen as the most prevalent in the prediction of the student's final grade. The two data are the number of materials downloaded and the number of times online. Therefore, these hypotheses are proposed:

- **H3:** Total number of materials downloaded has a direct correlation in the prediction of a student's final marks or grade.
- **H4:** Total number of times online has a direct correlation in the prediction of a student's final marks or grade.

Procedures

The data from classroom environment were directly taken from the lecturer of ABAP Course. The data were the pre-processed in order to make it applicable to be used with the data mining technique, multiple linear regression.

The target respondents for the questionnaire were the students who took ABAP course in the January 2011 semester. The questionnaire was randomly distributed (random sampling) to the students. The questionnaire was answered in a classroom environment while the students were having a break from their lecture session. It is believed that the respondents reveal an accurate representation of the course. All the data received towards the end of the process were then analyzed manually. The questionnaire was designed in a way, all the qualitative data can be represented by numerical values to make it appropriate for the data mining process. In order to do this, a scale of 1 to 5 is used to represent the answer for each question. From the data pre-processing, two questions were left out due to the fact that it may not be relevant enough to be used for this study. So, there were only 10 questions left. Besides that,

the questionnaire is also designed in which, the higher the result of the survey, the higher the final marks of the student should be.

In order to collect the E-Learning tracking data, a permission to access ABAP course module as the administrator was requested and granted. After a thorough observation on the available tracking data that are available in the old E-Learning system, it is decided that the data that will be extracted from it are the number of materials downloaded and the number of times online for each student. Sample of the data are as shown below:

Time	IP Address	Full name	Action	Information
Tue 1 February 2011, 02:47 PM	160.0.56.132	Khairulanwar B Mohd Noh 11793	forum user report	Muhammad Nasiruddin Bin Mohd Kamal 12752
Tue 1 February 2011, 02:47 PM	160.0.56.132	Khairulanwar B Mohd Noh 11793	user view	Muhammad Afiiq Bin Mohamed Ibrahim 12714
Tue 1 February 2011, 02:47 PM	160.0.56.132	Khairulanwar B Mohd Noh 11793	user view	Muhammad Nasiruddin Bin Mohd Kamal 12752
Tue 1 February 2011, 02:46 PM	160.0.56.132	Khairulanwar B Mohd Noh 11793	course view	Bussiness Application Programming
Time	IP Address	Full name	Action	Information
Tue 8 February 2011, 02:46 PM	160.0.56.99	Khairulanwar B Mohd Noh 11793	resource view	Lecture 1
Tue 8 February 2011, 02:46 PM	160.0.56.99	Khairulanwar B Mohd Noh 11793	resource view	Lecture 1
Tue 8 February 2011, 02:46 PM	160.0.56.99	Khairulanwar B Mohd Noh 11793	resource view all	
Tue 8 February 2011, 02:45 PM	160.0.56.99	Khairulanwar B Mohd Noh 11793	course view	Bussiness Application Programming
Tue 8 February 2011, 02:45 PM	160.0.56.99	Khairulanwar B Mohd Noh 11793	course view	Bussiness Application Programming

Figure 7: Sample of CMS Tracking Data

Resource view (in Action column) shows the number of materials downloaded for the student; meanwhile course view (in Action column) shows the student's number of times online. The E-learning tracking data collected is from January 2011 (starting from the start of the semester, which is on the 24th of January) until May 2011 (end of semester after final exam, which is until the 22nd of May 2011). All the data were counted manually because the old E-Learning system was unable to generate reports and statistics due to the fact that the system server and databases have already been inactivated. All the data were gathered for each student taking ABAP Course in Semester January 2011, for each day from 24th of January 2011 until 22nd of May 2011.

Deployment of data mining technique

In deciding the data mining techniques to be used in this project, a thorough study was done to choose the most appropriate and applicable technique in order to generate the expected result of predicting the students' final grade. Multiple linear regression was then chosen to be used in this project. Multiple linear regression is an extension of linear regression involving more than one predictor variable. Since there are four variables that are being studied to predict the final grade of the students, and the fact that each variable should have a direct relationship with the expected result, multiple linear regression technique is seen as the best technique to be deployed. Below is the least-squares multiple linear regression formula that is being used to calculate the prediction of the student's final marks

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

where,

a = the Y intercept / regression constant

b_1 = partial regression coefficient of the first independent variable (X_1) on Y

b_2 = partial regression coefficient of the second independent variable (X_2) on Y

b_3 = partial regression coefficient of the third independent variable (X_3) on Y

b_4 = partial regression coefficient of the fourth independent variable (X_4) on Y

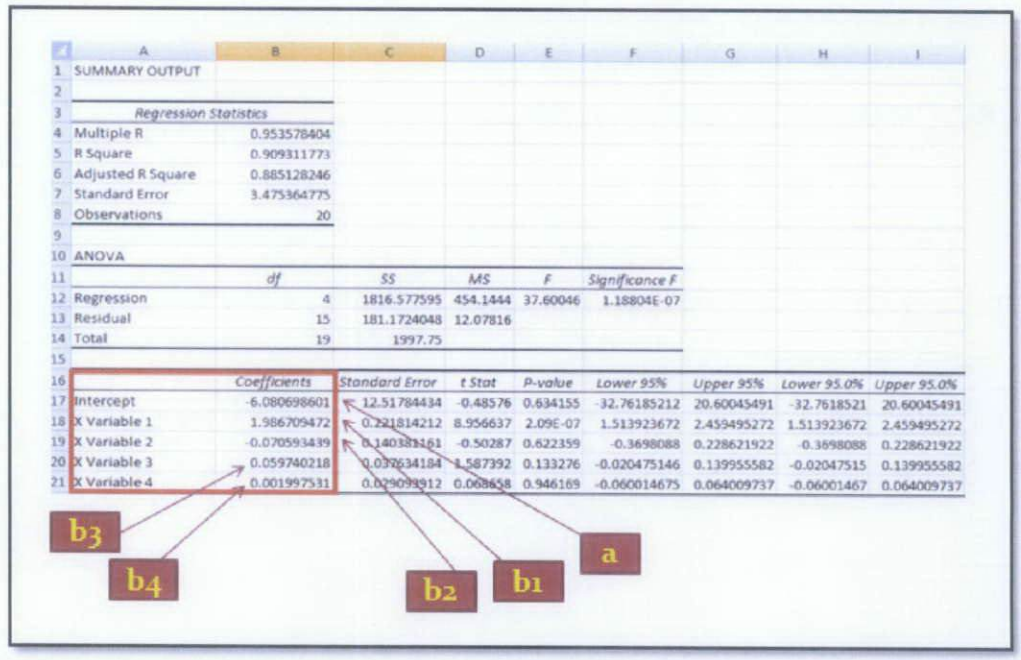
X_1 = First independent variable; *Coursework marks*

X_2 = Second independent variable; *Questionnaire result*

X_3 = Third independent variable; *E-Learning Data: Total number of materials downloaded*

X_4 = Fourth independent variable; *E-Learning Data: Total number of times online*

Partial regression coefficient (b_1, b_2, b_3, b_4) gives the amount by which the dependent variable (Y) increases when one independent variable (X_n) is increased by one unit and all the other independent variables are held constant. This coefficient is called partial because its value depends on the other independent variables exist in the formula. Due to its complexity and time consuming process to calculate and obtain the values, the use of software tools to is highly recommended. In this case, a Microsoft Excel software tool is used to obtain the value of each partial regression coefficient (b_1, b_2, b_3, b_4). A sample of the result is depicted in the figure below:



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.953578404							
5	R Square	0.909311773							
6	Adjusted R Square	0.885128246							
7	Standard Error	3.475364775							
8	Observations	20							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	4	1816.577595	454.1444	37.60046	1.18804E-07			
13	Residual	15	181.1724048	12.07816					
14	Total	19	1997.75						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	-6.080698601	12.51784434	-0.48576	0.634155	-32.76185212	20.60045491	-32.7618521	20.60045491
18	X Variable 1	1.986709472	0.221814212	8.956637	2.09E-07	1.513923672	2.459495272	1.513923672	2.459495272
19	X Variable 2	-0.070593439	0.140381161	-0.50287	0.622359	-0.3698088	0.228621922	-0.3698088	0.228621922
20	X Variable 3	0.059740218	0.037634184	1.587392	0.133276	-0.020475146	0.139955582	-0.02047515	0.139955582
21	X Variable 4	0.001997531	0.029099912	0.068658	0.946169	-0.060014675	0.064009737	-0.06001467	0.064009737

Figure 8: Partial Regression Coefficient

CHAPTER 4

RESULTS AND DISCUSSION

Results and discussion

Demographics: A total number of 28 respondents responded to the questionnaire. Data from 6 respondents were incomplete so discarded. Hence, the remaining 22 respondents were used for the analyses. Hypotheses testing were done using an online Statistical Calculator for Correlation Coefficient, Coefficient of determination and Scatter Plot Generator. Regression analysis was taken to test the hypothesis of this study as this technique is one of the most common and appropriate method for hypothesis testing [26]. The results of the hypotheses testing are as depicted in the diagrams and explained below:

H1: Value of 'R' for hypothesis H1 is 0.9355. This means that strong positive correlation exists between coursework marks and the student's final grade. R-squared value is 0.8751 shows that if one unit of independent variable increases then 0.8751 units of dependent variable will increase. The closer the R-Squared value, the closer the value of another term can be predicted using the value of one term. As depicted in the scatter plot diagram on the next page, there is a strong positive relationship between coursework marks and student's final grade.

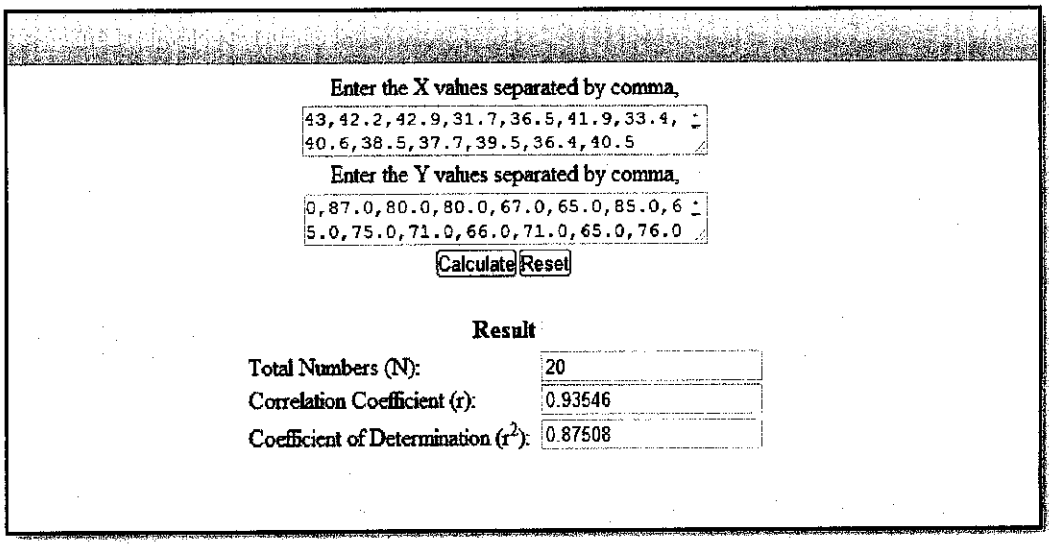


Figure 9: Correlation Coefficient and Coefficient of determinant for variable Y and X1

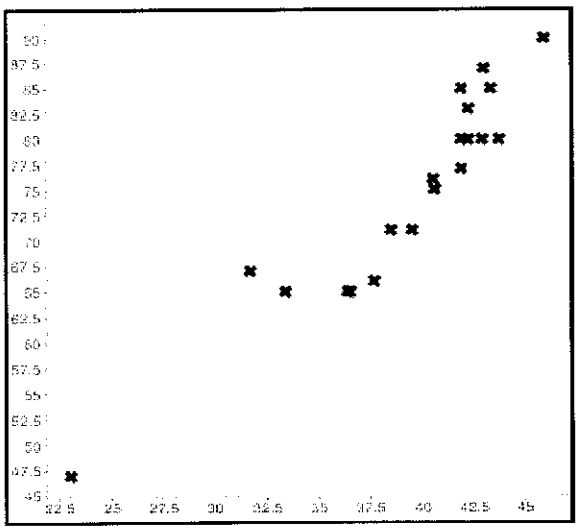


Figure 10: Scatter Plot for variable Y and X1

H2: Value of 'R' for hypothesis H2 is -0.1375. This means that weak negative correlation exists between survey results and student's final grade. R-Squared value is 0.0189, which can be considered as weak. More generally, a lower value of R-Squared means that you might not be able to predict one term from another accurately.

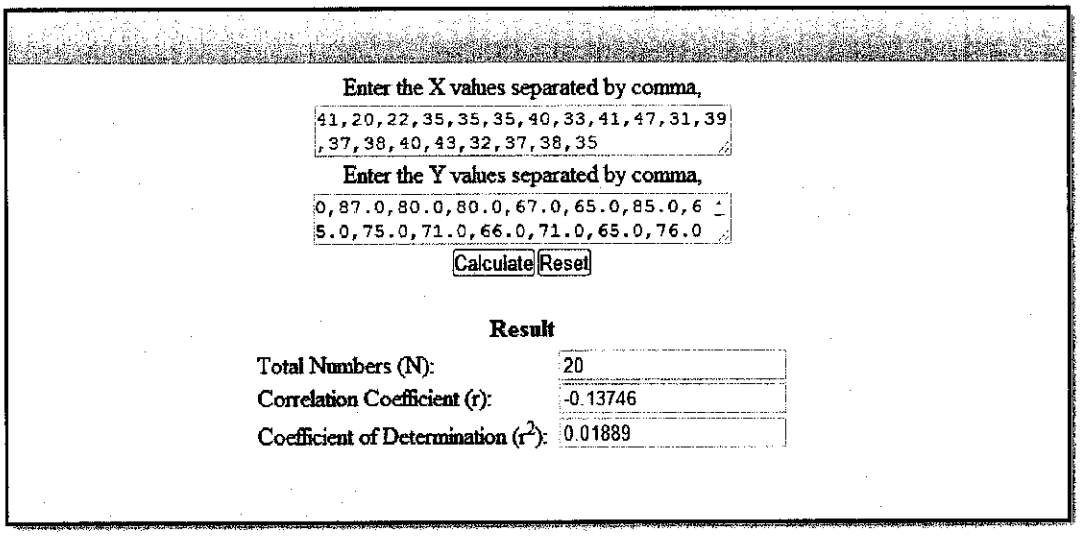


Figure 11: Correlation Coefficient and Coefficient of determinant for variable Y and X2

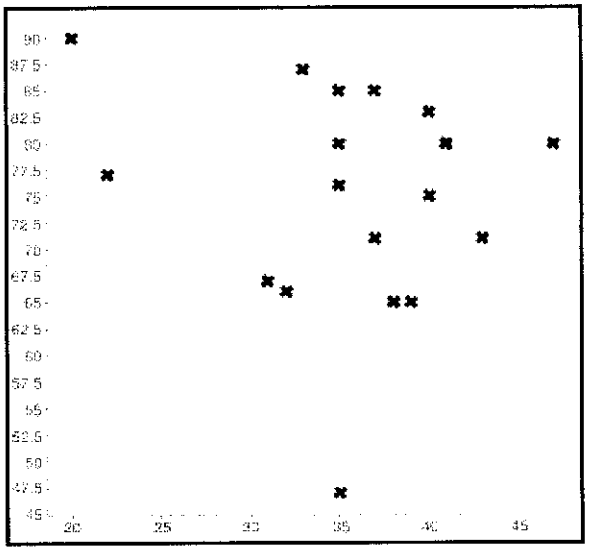


Figure 12: Scatter Plot for variable Y and X2

H3: Value of 'R' for hypothesis H3 is -0.2731. This means that weak negative correlation exists between the number of materials downloaded and student's final grade. R-Squared value is 0.0746, which can also be considered as weak.

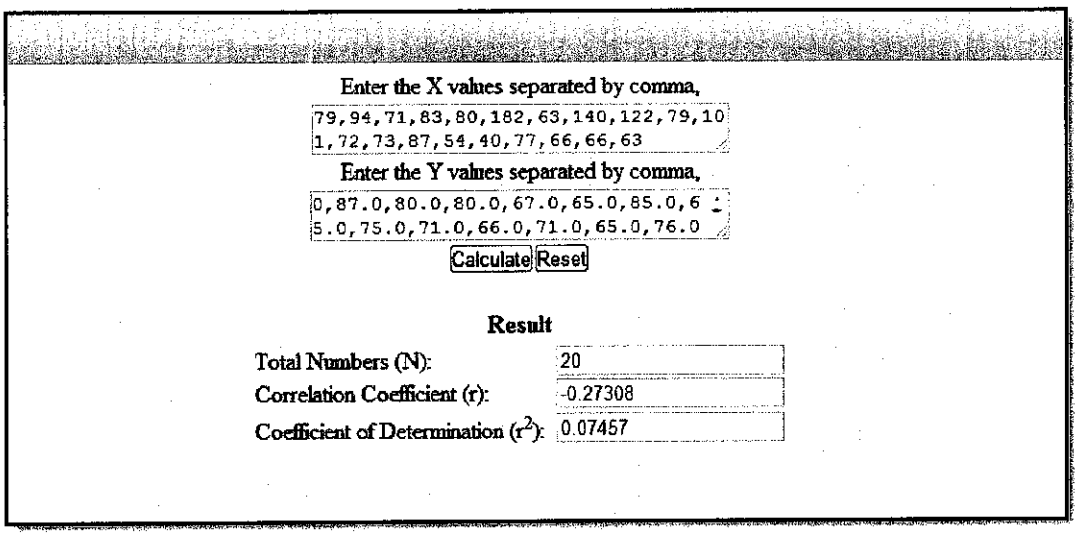


Figure 13: Correlation Coefficient and Coefficient of determinant for variable Y and X3

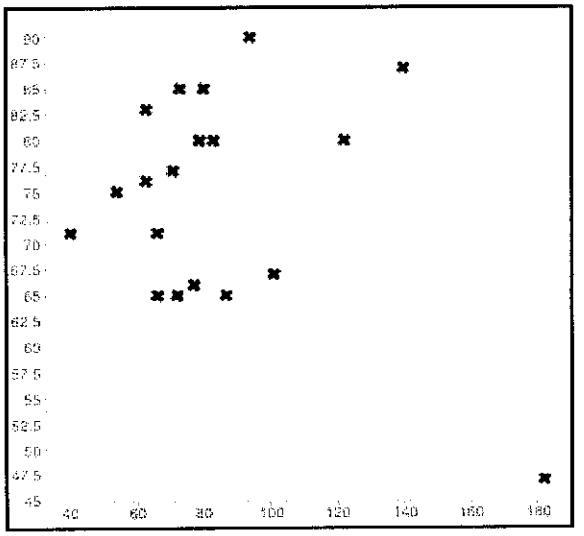


Figure 14: Scatter Plot for variable Y and X3

H4: Value of 'R' for hypothesis H4 is 0.3822. This means that weak positive correlation exists between the number of times online and the student's final grade. R-Squared value is 0.1461 shows that if one unit of independent variable increases then 0.1461 units of dependent variable will increase.

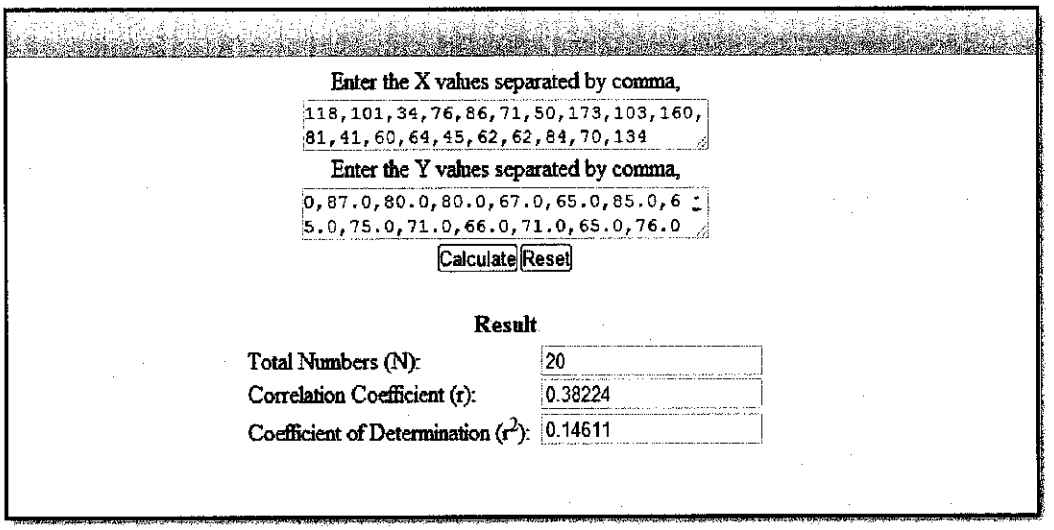


Figure 15: Correlation Coefficient and Coefficient of determinant for variable Y and X4

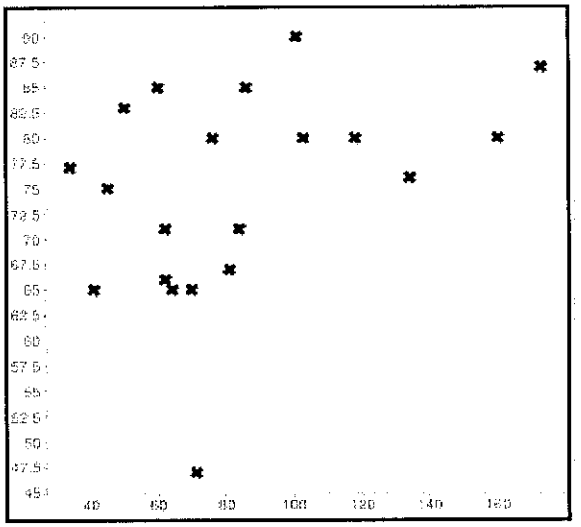


Figure 16: Scatter Plot for variable Y and X4

Snapshots of the system

The snapshots of the system are as shown in the figures below. The design of the system is subject to further enhancement and modifications to better suit the user.

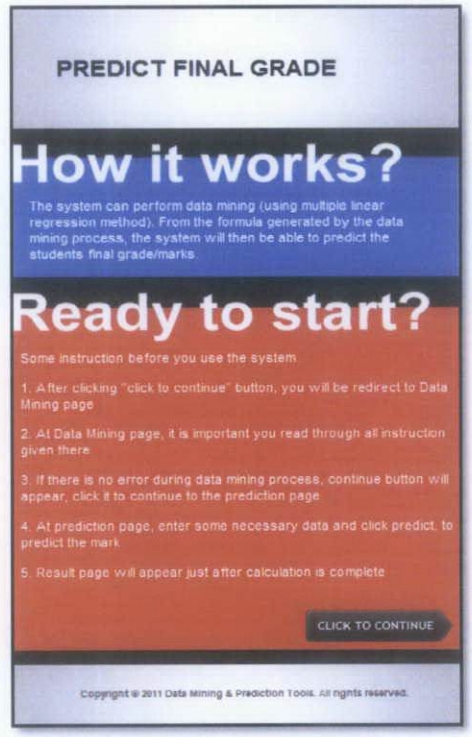
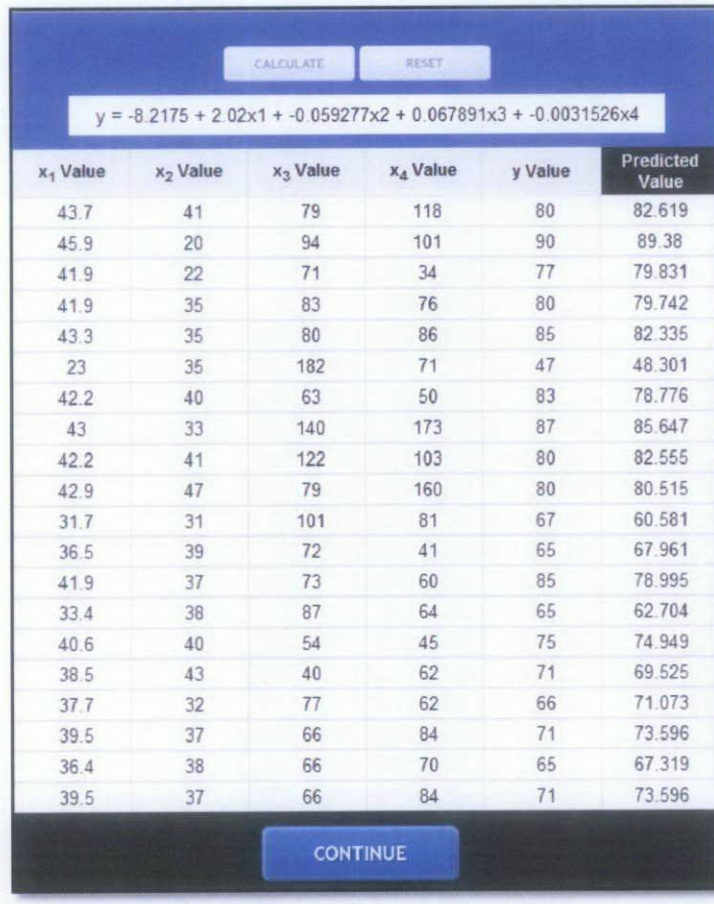


Figure 17: Start Page

Page Description: This page explains the preliminary information about the system, such as how the system works and some instructions on how to use the system.



x_1 Value	x_2 Value	x_3 Value	x_4 Value	y Value	Predicted Value
43.7	41	79	118	80	82.619
45.9	20	94	101	90	89.38
41.9	22	71	34	77	79.831
41.9	35	83	76	80	79.742
43.3	35	80	86	85	82.335
23	35	182	71	47	48.301
42.2	40	63	50	83	78.776
43	33	140	173	87	85.647
42.2	41	122	103	80	82.555
42.9	47	79	160	80	80.515
31.7	31	101	81	67	60.581
36.5	39	72	41	65	67.961
41.9	37	73	60	85	78.995
33.4	38	87	64	65	62.704
40.6	40	54	45	75	74.949
38.5	43	40	62	71	69.525
37.7	32	77	62	66	71.073
39.5	37	66	84	71	73.596
36.4	38	66	70	65	67.319
39.5	37	66	84	71	73.596

Figure 18: Data Mining Page

Page Description: The purpose of this page is to build the model (using multiple linear regression) to be used in the prediction of the student's final grade. The user will have to enter some data before the system can generate a regression formula. In this study, X1 represents the variable coursework marks, X2 represents the variable questionnaire result, X3 represents the variable total number of materials downloaded (CMS Tracking Data), X4 represents the variable total number of times online (CMS Tracking Data) and Y represents the actual final exam marks of the students. The column predicted value is used is to display the predicted Y values. These values and the regression formula will be generated once the user press the calculate button.

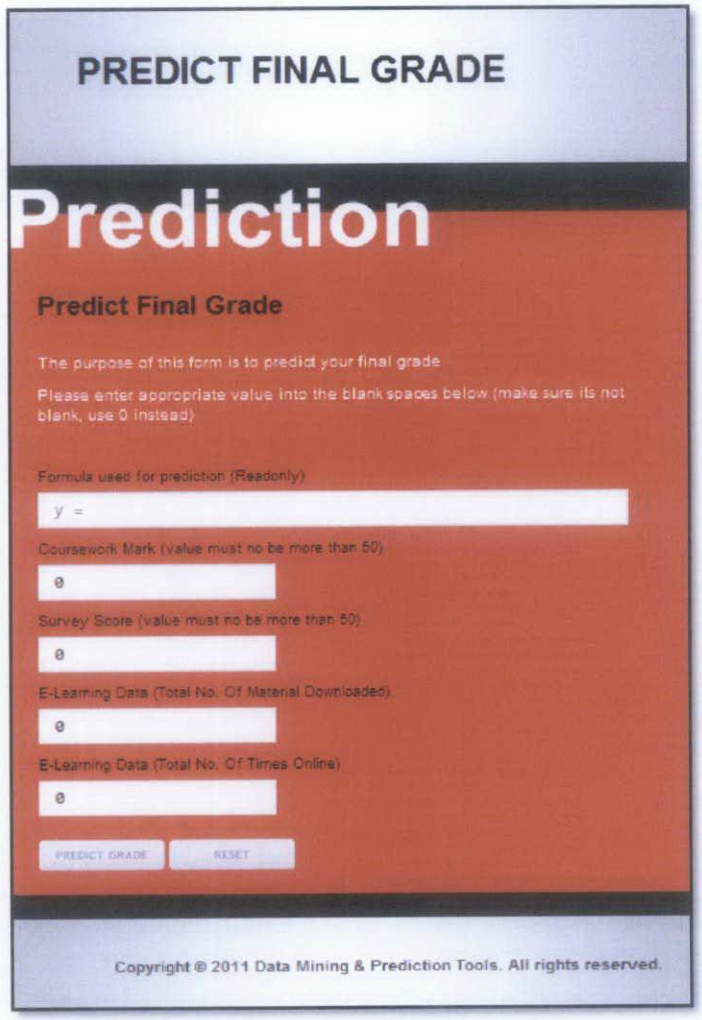


Figure 19: Prediction Page

Page Description: This page is used as a form to obtain the necessary inputs from the user in order to make a prediction. The formula generated from the data mining page previously will be transferred to the read-only text field and will be used in the calculation to make the prediction. User will have to key in the data such as coursework marks, survey score (or questionnaire result), total number of materials downloaded and total number of times online.

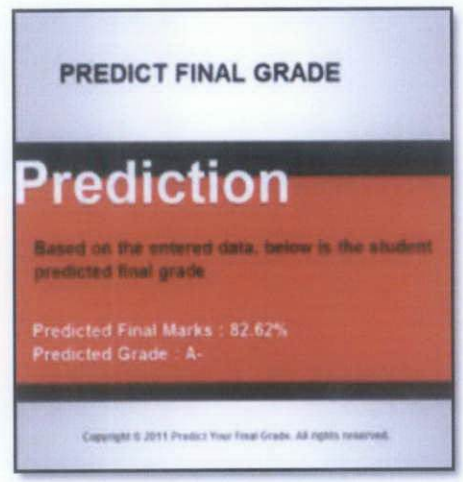


Figure 20: Prediction Result Page

Page Description: This purpose of this page is to display the prediction result generated by the system.

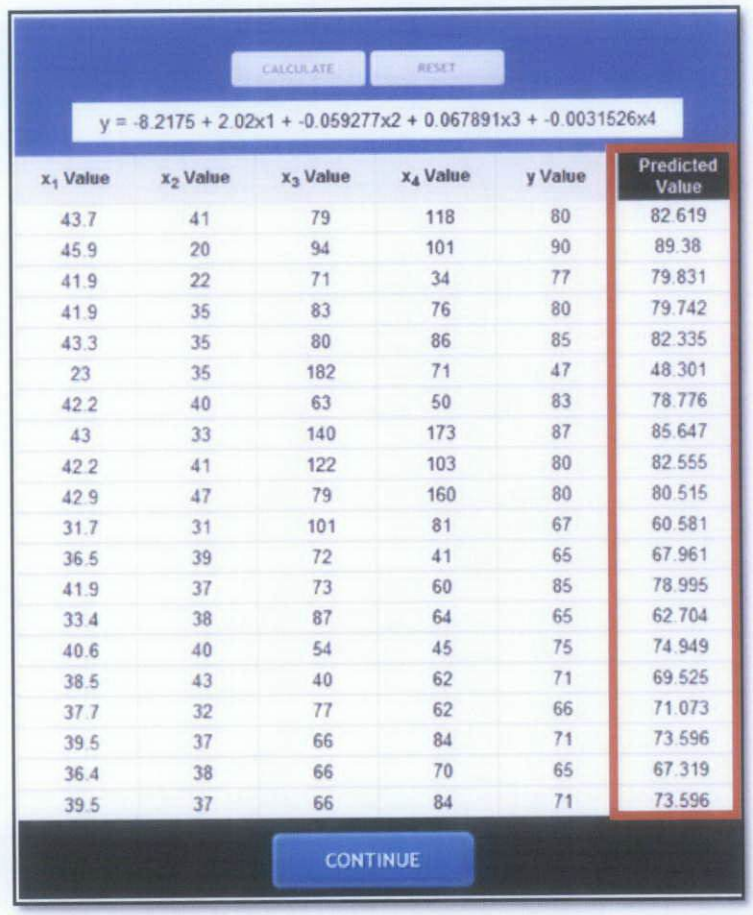
Functionality Test

The table below shows the functionality test that was done to the system:

Case	Functionality Tested	Input Details	Expected Result	Actual Result
1	Start Page	Continue Button	Preliminary info about the system. Continue button redirect user to data mining page.	Successful
2	Module 1 – Data mining page	Calculate Button, Reset Button, Continue Button	Model building – Perform data mining using multiple linear regression and generate predicted values and regression formula when user click calculate button. Continue button redirect user to prediction page.	Successful
3	Module 2 – Prediction Page	Text fields, Predict Grade Button, Reset Button	Display formula generated from data mining page. User can input data in this page for prediction purpose. Predict grade button redirect user to the prediction result page.	Successful
4	Module 2 – Prediction Result Page	N/A	Display student's predicted final marks and grade.	Successful

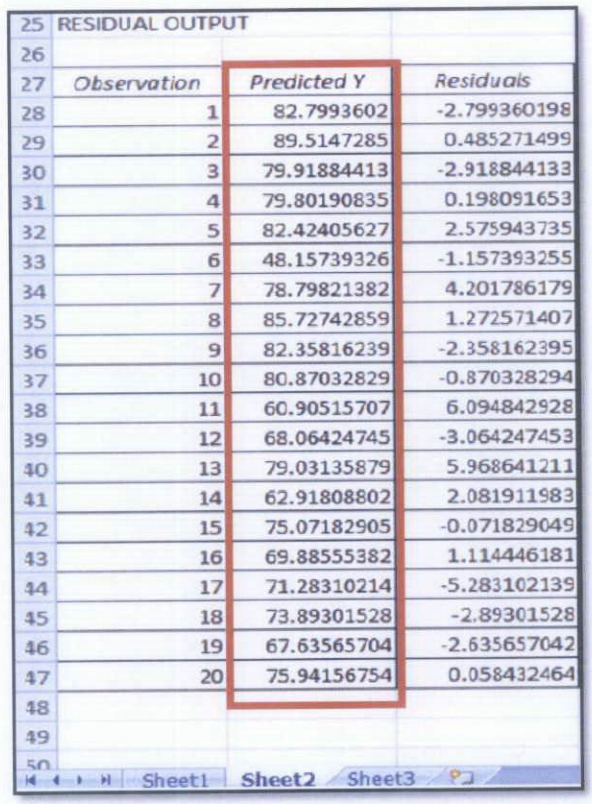
Table 2: Functionality Test

In order to compare the accuracy of the model used in the system, the results generated by the system were then compared with another commercialized system which is Microsoft Excel Regression Analysis Tool. As we can see from the figures below and in the next page, the results generated for both systems are almost similar with one another.



x_1 Value	x_2 Value	x_3 Value	x_4 Value	y Value	Predicted Value
43.7	41	79	118	80	82.619
45.9	20	94	101	90	89.38
41.9	22	71	34	77	79.831
41.9	35	83	76	80	79.742
43.3	35	80	86	85	82.335
23	35	182	71	47	48.301
42.2	40	63	50	83	78.776
43	33	140	173	87	85.647
42.2	41	122	103	80	82.555
42.9	47	79	160	80	80.515
31.7	31	101	81	67	60.581
36.5	39	72	41	65	67.961
41.9	37	73	60	85	78.995
33.4	38	87	64	65	62.704
40.6	40	54	45	75	74.949
38.5	43	40	62	71	69.525
37.7	32	77	62	66	71.073
39.5	37	66	84	71	73.596
36.4	38	66	70	65	67.319
39.5	37	66	84	71	73.596

Figure 21: Results generated by the system



25	RESIDUAL OUTPUT		
26			
27	*Observation*	*Predicted Y*	*Residuals*
28	1	82.7993602	-2.799360198
29	2	89.5147285	0.485271499
30	3	79.91884413	-2.918844133
31	4	79.80190835	0.198091653
32	5	82.42405627	2.575943735
33	6	48.15739326	-1.157393255
34	7	78.79821382	4.201786179
35	8	85.72742859	1.272571407
36	9	82.35816239	-2.358162395
37	10	80.87032829	-0.870328294
38	11	60.90515707	6.094842928
39	12	68.06424745	-3.064247453
40	13	79.03135879	5.968641211
41	14	62.91808802	2.081911983
42	15	75.07182905	-0.071829049
43	16	69.88555382	1.114446181
44	17	71.28310214	-5.283102139
45	18	73.89301528	-2.89301528
46	19	67.63565704	-2.635657042
47	20	75.94156754	0.058432464
48			
49			
50			

Sheet1 Sheet2 Sheet3

Figure 22: Results generated by Microsoft Excel Regression Analysis Tools

Besides that, the average percentage of error for all the values from generated from the system were calculated. Percentage of error is the difference between the Approximate and Exact Values, as a percentage of the Exact Value. The formula to calculate percentage of error is as shown below:

$$\frac{|\text{Approximate Value} - \text{Exact Value}|}{|\text{Exact Value}|} \times 100\%$$

Actual Result	Result generated by the system	Percentage of Error (1)	Result generated by Excel	Percentage of Error (2)
80.0	82.619	3.274	82.799	0.217
90.0	89.38	6.889	89.515	0.151
77.0	79.831	3.677	79.919	0.110
80.0	79.742	0.323	79.802	0.075
85.0	82.335	3.135	82.424	0.108
47.0	48.301	2.768	48.157	0.299
83.0	78.776	5.089	78.798	0.028
87.0	85.647	1.555	85.727	0.093
80.0	82.555	3.194	82.358	0.239
80.0	80.515	0.644	80.870	0.439
67.0	60.581	9.581	60.905	0.532
65.0	67.961	4.555	68.064	0.151
85.0	78.995	7.065	79.031	0.046
65.0	62.704	3.532	62.918	0.340
75.0	74.949	0.068	75.072	0.164
71.0	69.525	2.077	69.886	0.516
66.0	71.073	7.686	71.283	0.295
71.0	73.596	3.656	73.893	0.402
65.0	67.319	3.568	67.636	0.469
76.0	73.596	3.163	75.942	3.089
Average Percentage of Error		**3.775**		**0.388**

Table 3: Percentages and Average Percentages of Error

Percentage of Error (1) is the comparison between the result generated from the system (Approximate Value) with the actual result or final marks of the students (Exact Value). The average percentage of error between the two subjects is 3.775%.

Percentage of Error (2) is the comparison between the results generated from the system (Approximate Value) with the result generated using another commercialized system; Microsoft Excel Regression Analysis tools (Exact Value). The average percentages of error between the two subjects are 0.388%.

Problems and Challenges

There are a few problems and challenges faced throughout the process of completing this final year project. One of the main challenges is to find during the analysis phase, in which, the most suitable data mining technique to be used needs to be decided. With limited knowledge in Data Mining, a thorough study had to be done in order to choose the best technique. Besides that, the process of extracting data from the existing CMS in the university was also challenging. This is due to the fact that the data that needs to be extracted are historical data and the CMS in the university has just been migrated to a new system. Having said that, the server for the old system has been deactivated and for that reason the old CMS is unable to generate reports automatically with regard to the online activities for each student. The process was tedious as all the data were extracted manually.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

Conclusion

This project highlights one of the most important research areas in the modern educational system which is Educational Data Mining. The design and development of the tools will be based from one of the various data mining techniques which will be used to analyze different kind of data coming from different educational settings and predict students' performance in a programming course. The main contribution of this project is the development of a new data mining and prediction tools to support the teaching process of lecturer of a particular programming course.

This study also provides an empirically tested theoretical model that shows how data from different educational settings can contribute in the prediction of student's final grade. The results indicate that coursework marks has the most significant positive relationship with the student's final grade followed by one of the CMS tracking data which is the total number of materials downloaded. The other two variables, which are the questionnaire on psychosocial factors and the total number of materials downloaded that were studied however were found to have a weak negative relationship with the student's final grade.

Recommendations

Future work should be focusing on the integration of this tool into the existing Course Management System (CMS) in the university. This will produce a more reliable system, in which, it does not only supports the day-to-day teaching and learning process, but is also able to perform analysis based on many kinds of data extracted from the system itself as well as external data that can be stored into the system (e.g. students' attendance records, test marks etc). Besides that, the integration of this tool with the existing CMS in the university will automate the data input process, in which the CMS tracking data can be extracted directly from the

existing system. Apart from that, the new development of data mining and prediction tools may also be implemented for other courses and not just the programming courses, as per in the context of this project. Lastly, future work conducted in this scope should provide more analysis and interpretation on other variables that might have the impact on student's final grade.

References

[1] *Educational Data Mining.* Retrieved March 4, 2011 from Educational Data Mining website <http://www.educationaldatamining.org/>

[2] Romero, C. et al., Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

[3] Macfadyen L.P., Dawson S., Mining LMS data to develop an "early warning system" for educators: A proof of concept, *Computers & Education* (2010), 588-599

[4] Hung J.L., Zhang K., Revealing Online Learning Behaviours and Activity Patterns and Making Predictions with Data Mining Techniques in Online Learning (2006)

[5] White B., Larusson J., Seeing, Thinking, Doing: Strategic Directive for LMS (2009)

[6] Mohamad Farhan Mohamad Mohsin, Mohd Helmy Abd Wahab, Mohd Fairuz Zaiyadi, Morita Md Norwawi, Cik Fazilah Hibadullah, An Investigation into Influence Factor of Student Programming Grade Using Association Rule Mining, *Advances in Information Sciences and Service Sciences Volume 2, Number 2* (2010), doi: 10.4156/aiss.vol2.issue2.3

[7] Kinnunen P., Murphy L., McCartney R., Thomas L., Through the eyes of instructors: a phenomenographic investigation of student success, *K.3.2 [Computers & Education]: Computer & Information Science Education – Computer Science Education* (2007)

[8] Norhidayah Ali, Kamarulzaman Jusoff, Syukriah Ali, Najah Mokhtar, Azni Syafena Andin Salamat, The Factors Influencing Students' Performance at Universiti Teknologi MARA, Kedah, Malaysia, *Management Science and Engineering ISSN 1913-0341,* Vol.3 No.4 (2009)

[9] Womble L.P., Impact of stress factors on college student academic performance (2004)

[10] *Class Attendance Article.* Retrieved March 2, 2011 from mnsu website < http://www.mnsu.edu/cetl/teachingresources/articles/classattendance.html>

[11] Fraser W.J., Killen R., Factors influencing academic success or failure of first-year and senior university students: do education students and lecturers perceive things differently?, *South African Journal of Education,* Vol 23(4) 254 – 260 (2003)

[12] *Psychosocial.* Retrieved March 2, 2011 from Wikipedia website <http://en.wikipedia.org/wiki/Psychosocial>

[13] Mongan-Rallis H., Jugovich S.M., Comparison of WebCT & Moodle Course Management Systems (CMS), *Presentation to M.Ed. Program* (2007)

[14] Data Mining Techniques, Retrieved March 2, 2011 from StatSoft website < http://www.statsoft.com/textbook/data-mining-techniques/>

[15] Dennis A., Wixom B.H., Tegarden D. (2005). System Analysis and Design, 2nd edn. Massachusetts: John Wiley & Sons.

[16] Data: Mining with a Mission, Retrieved March 3, 2011 from tech Learning website <http://www.techlearning.com/story/showArticle.jhtml?articleID=18311595>

[17] Baker R.S.J.D., Yacef K., The State of Educational Data Mining in 2009: A Review and Future Visions (2009)

[18] Bergin S., Reilly R., Programming: Factors that Influence Success, *Department of Computer Science* (2005)

[19] Bergin S., Reilly R., The influence of motivation and comfort-level on learning to program, *Department or Computer Science,* Pages 293-304 (2005)

[20] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M.Inayat Khan, Data Mining Model for Higher Education System, *European Journal of Scientific Research,* Vol.43 No.1, pp.24-29 (2010)

[21] Merceron A., Yacef K., Educational Data Mining: a Case Study, *School of Information Technologies – University of Sydney, Australia* (2005)

[22] White G., Sivitanides M., An Empirical Investigation of the Relationship Between Success in Mathematics and Visual Programming Courses, *Journal of Information Systems Education,* Vol. 14(4) (2005)

[23] Ogor E.N., Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques, *Department of Natural Sciences,* doi 10.1109/CERMA.2007.78 (2007)

[24] Roiger R.J. & Geatz M.W. (2003). *Data Mining A Tutorial-Based Primer.* Addison-Wesley, Pearson Education.

[25] Han J. & Kamber M. (2001). *Data Mining Concepts and Techniques.* Academic Press, Morgan Kaufmann.

[26] Jehangir M., Dominic P.D.D., Downe A.G., Business Resources Impact on E-Commerce Capability and E-Commerce Value: An Empirical Investigation, *Department of Computer and Information Sciences, Universiti Teknologi Petronas ISSN 1819-3579, DOI: 10.3923/tasr.2011* (2011)

Appendix 1: Gantt Chart

Appendix 1

FYP 2 - Project Timeline

Title: Data Mining and Prediction Tools: for Predicting Students' Performance in Programming Course

ID	Task Name	Duration (week)	Start	Finish	May '11		June '11				Jul '11				Aug '11			
					Week 1 (23/5)	Week 2 (30/5)	Week 3 (6/6)	Week 4 (13/6)	Week 5 (20/6)	Week 6 (27/6)	Week 7 (4/7)	Week 8 (11/7)	Week 9 (18/7)	Week 10 (25/7)	Week 11 (1/8)	Week 12 (8/8)	Week 13 (15/8)	Week 14 (22/8)
1	Progress Report	6	23-May	8-Jul														
2	Pre-EDX	2	11-Jul	1-Aug														
3	Dissertation	1	1-Aug	9-Aug														
4	VIVA	2	9-Aug	19-Aug														
4	Final Dissertation	2	19-Aug	26-Aug														

Legend:	
	Start Date and Duration
	Deadline

Prepared By: Che Sarah Che Nordin
Student ID: 11258

Appendix 2: Interview Notes

Person Interviewed: Ms. Norkamar Faridatul Salwa Bt. Kamarudin
(Senior Executive, IT Multimedia Services Department, Universiti Teknologi Petronas)

Interviewer: Che Sarah Che Nordin
Date: 22nd of March 2011 (Tuesday)

1. **Purpose of Interview:**

- Studying the background of UTP E-learning system (as-is system).
- Determine what kind of information can be extracted from the system's server logs.

2. **Summary of the interview:**

- UTP E-learning system is using Moodle v1.6 online learning software package.
- Since the system is open-source, the administrator has the right to modify or reprogram the system. However, this is rarely been done and only some modifications were done previously just to standardized the design of the system.
- The person who has the highest level is the system administrator, which she can also all the course modules in the university. The lecturer also has the administrator access but only limited to the course in which he or she is teaching.
- The history or records of past students' grades (grades that were captured online), will be there in the system, if the lecturer still save everything. So it depends on the lecturer themselves whether they want to keep the data or not.
- For assignments, tests or examinations that were done using normal paper, the system won't be able to capture and store these data. Therefore, no data mining can be done for these data. These data can only be stored in Student

Prism Portal, which acts more like the university's Student Management System.

- Future plan: to integrate the UTP E-learning together with Student Prism Portal under one platform; standardization.
- The system stores quite a number of information on students' profile but not as much as Student Prism Portal.
- The system is using SQL database and PHP as its programming language. However, it is not so stable since there are a vast amount of transactions happening everyday and the system only has one server (and issue of database overloading) to cater to this activity. Another two servers will be added to backup the system.
- In UTP, no expert users. In other universities, it is collaboration between the IT department and the lecturers to support the daily operation of the system.

Appendix 3: Questionnaire on Psychosocial Factors

Final Year Project Survey

Analysis & Data Mining Tools

(for predicting student success in programming course)

Based on your previous experience in taking the programming course, Advanced Business Application Programming (ABAP), in order to obtain a good grade in programming subject, do you think these factors or values are important. Please indicate your learning outcomes (level of ability and satisfaction) in each of the factor stated below.

Student ID: _____

Factor	Low				High
Understanding about the concepts and topics taught in the programming subject	1	2	3	4	5
Interest in programming subject	1	2	3	4	5
Willingness to accept challenge & desire to learn, high self-motivation and discipline in programming course	1	2	3	4	5
Willingness to ask for help from lecturer or tutor when encountered with problems	1	2	3	4	5
Consistent efforts to learn programming	1	2	3	4	5
Well-structured presentations by lecturer	1	2	3	4	5
Clear understanding of the lecturer's expectation on assignments, projects, tests and examinations	1	2	3	4	5
Encouragement, motivation and support from lecturer in understanding the concepts and topics	1	2	3	4	5
Availability of quality learning	1	2	3	4	5

resources (E.g. books in library, materials provided by lecturer)					
Satisfactory accommodations (E.g. Computer labs and good hostel conditions)	1	2	3	4	5
Positive influence and support by friends, coursemates and peer groups.	1	2	3	4	5
Stable financial status	1	2	3	4	5

Thank you very much for participating in this survey! Your time spent is highly appreciated! ☺